Computational speech segregation inspired by principles of auditory processing

Bentsen, Thomas

Publication date:
2018

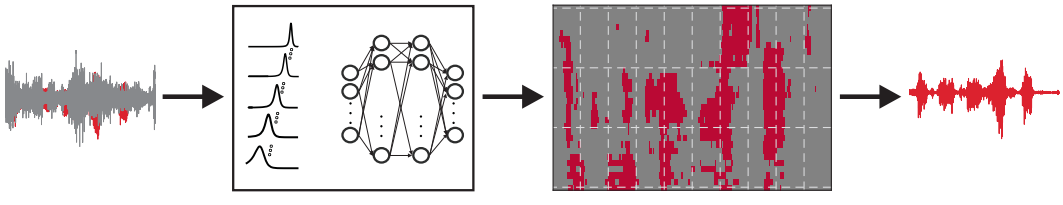
Document Version
Publisher's PDF, also known as Version of record

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 33

Thomas Bentsen

Computational speech segregation inspired by principles of auditory processing



Hearing Systems
Department of Electrical Engineering

Computational speech segregation inspired by principles of auditory processing

PhD thesis by
Thomas Bentsen

Technical University of Denmark

2018

Cover illustration by Thomas Bentsen.
The thesis was defended on 5 February 2018.

This PhD thesis is the result of a research project within the Hearing Systems group, Department of Electrical Engineering, Technical University of Denmark.

The project was partly financed by the Oticon Centre of Excellence for Hearing and Speech Sciences (2/3) and partly financed by the Technical University of Denmark (1/3).

The assessment committee consisted of Associate Professor Jeremy Marozeau, Professor Stefan Bleeck, and Professor Jürgen Tchorz.

Supervisors

Assistant Professor Tobias May
Postdoc Abigail Anne Kressner
Professor Torsten Dau
Hearing Systems Group
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Abstract

Understanding speech in noise in adverse listening conditions can be challenging for many people, in particular hearing-aid users and cochlear-implant recipients. To improve the speech understanding, better noise reduction strategies are needed in such devices. The performance of the strategies depends on how well the characteristics of the speech and the noise are known. Therefore, it is necessary to have automatic approaches that can separate the speech from the noise as accurate as possible, which is the overall goal of computational speech segregation. Often, an ideal time-frequency mask is estimated in these approaches. In the mask, the level of speech activity is indicated in each time-frequency unit. The mask is estimated by extracting auditory-inspired features from the noisy speech and subsequently learning the characteristics of the speech and noise with machine-learning techniques. This thesis investigated three approaches within computational speech segregation based on ideal time-frequency mask estimation. The approaches were evaluated in the framework of noise reduction to improve speech understanding of normal-hearing listeners and cochlear-implant recipients in noisy environments.

In the first approach, machine-learning techniques were employed in separate auditory frequency bands to classifying each mask unit as either speech-dominated or noise-dominated. Words are composed of phonemes that may occupy several neighboring units in the estimated mask. The focus was on how to use this contextual information in speech across time and frequency in computational speech segregation. Exploiting the context across frequency was found to be important. By increasing the amount of considered spectral information, higher measured speech intelligibility was obtained in normal-hearing listeners. On the other hand, exploiting the context across time in computational speech segregation is perhaps not a critical factor to increase speech intelligibility. Recent approaches within computational speech

segregation are based on deep neural networks, and speech intelligibility improvements have successfully been demonstrated in adverse conditions. In a second approach, a deep neural network was therefore employed and the roles and the relative contribution of a selection of components, that may be responsible for the success, were analyzed. Two components, namely the network architecture and the estimation of an ideal time-frequency mask based on continuous gain values, were found to play a significant role. In a third approach, an application of the estimated time-frequency mask was considered in real-time cochlear-implant processing. A proposed speech coding strategy selects cochlear-implant channels for electrical stimulation, and only if the signal-to-noise ratio within the channel is larger or equal to a local criterion. However, this strategy relies on ideal signal-to-noise ratios and a noise power estimation stage is, therefore, required to estimate the signal-to-noise ratios in real-time cochlear-implant processing. Results implied that a noise power estimation with improved noise-tracking capabilities does not necessarily translate to increased speech intelligibility. However, the adaptive channel selection is important for reducing the noise-induced stimulation in the cochlear-implant recipients.

Overall, the results of this thesis have implications for the design of computational speech segregation approaches with noise-reduction applications. Furthermore, the results may guide the development of a single cost function, which correlates with speech intelligibility, to assess and optimize the system performance.

Resumé

For mange mennesker, især høreapparat- og høreimplantatbrugere, kan det være en udfordring at forstå tale i støjfyldte omgivelser. For at forbedre taleforståeligheden, er det nødvendigt med bedre støjreduktions-strategier i disse apparater. Effektiviteten af strategierne er afhængig af, hvor godt talen og støjens karakteristika kendes. Derfor er det nødvendigt med automatiske metoder, der kan adskille talen fra støjen så nøjagtigt som muligt, hvilket er målet med *computational speech segregation*. I disse metoder estimeres ofte en ideel tids-frekvens-maske. I masken er niveauet af taleaktivitet angivet i hver tids-frekvens enhed. Masken estimeres ved at uddrage auditorisk inspireret features og efterfølgende lære talens og støjens karakteristika med brug af *machine-learning* teknikker. I denne afhandling undersøgtes tre metoder inden for *computational speech segregation*, baseret på estimering af ideelle tidsfrekvens-masker. Metoderne blev evalueret inden for rammerne af støjreduktion med henblik på at øge taleforståeligheden for normalthørende og høreimplantat-brugere i støjfyldte miljøer.

I den første metode blev *machine-learning* teknikker anvendt i separate auditoriske frekvensbånd for at klassificere hver enhed i masken som værende enten tale- eller støjdominante. Ord er komponeret af fonemer, der kan fylde flere naboliggende enheder i den estimerede maske. Fokus var på, hvordan den kontekstuelle information kan bruges over tid og frekvens i *computational speech segregation*. Informationen i talen over frekvens er vigtig at udforske. Ved at inkludere mere information i talen over frekvens blev en højere taleforståelighed målt hos normalthørende. På den anden side er den temporale information i talen måske ikke afgørende at udforske i *computational speech segregation* for at øge taleforståeligheden. De nyeste metoder indenfor computational speech segregation er baseret på dybe neurale netværk, og med disse metoder er taleforståeligheden succesfuldt blevet forbedret i udfordrende lytte miljøer. I den anden metode blev et dybt neuralt netværk derfor anvendt,

og rollerne samt det relative bidrag fra en række komponenter, der måske kan forklare succesen, blev analyseret. To komponenter, netværksarkitekturen og estimeringen af en ideel tids-frekvens maske baseret på kontinuerte forstærkningsværdier, spillede en afgørende rolle. I den tredje metode blev en anvendelse af den estimerede ideelle tids-frekvens maske betragtet i realtidsbehandling i høreimplantater. En foreslået talekodningsstrategi udvælger frekvensbånd i høreimplantatet med henblik på elektrisk stimulering, såfremt signal-støj-forholdet i et frekvensbånd er større end eller lig et kriterie. Dog er denne strategi baseret på ideelle signal-støj-forhold og en algoritme, der kan estimere støjens effekt, er derfor nødvendig for at være i stand til at estimere signal-støj-forholdene i realtidsbehandling i høreimplantater. Resultaterne indebærer, at en algoritme med en forbedret egenskab til at følge støjen ikke nødvendigvis medfører en øget taleforståelighed. Dog er den adaptive udvælgelse af frekvensbåndene i metoden vigtig med henblik på at reducere støj-induceret stimulering i høreimplantatbrugere.

Samlet set har resultaterne i denne afhandling implikationer for design af metoder inden for *computational speech segregation*, der har til formål at reducere den omgivende støj. Desuden kan resultaterne guide udviklingen af et enkelt objektivt mål, der korrelerer med taleforståeligheden, til at vurdere og optimere virkningen af metoderne.

Acknowledgments

First and foremost, I wish to thank my supervisors Tobias May, Abigail Anne Kressner and Torsten Dau for inspirational discussions as well as invaluable guidance and support throughout the years. Your constructive feedback has been much appreciated, and I am thankful for teaching me the importance of being a communicator. Tobias and Abbie, it has been a pleasure working with you guys. Torsten, a special thanks for giving me the opportunity to work in the group. I could not have asked for a better team.

I would also like to express my gratitude to Cochlear Limited for giving me the opportunity to visit their research lab in Melbourne. A special thanks to Stefan Mauger for his tremendous support during the research stay, and the warm hospitality he and his family has shown me. Also, thanks to Adam Hersbach for technical help and for showing me the best beer taps while in Melbourne.

I wish to thank all of my colleagues in Room 111 for their encouragement and help, and Caroline van Oosterhout for her indispensable administrative support. Moreover, a thanks to Marianna Vatti from the Eriksholm Research Centre for fruitful knowledge-sharing meetings.

To my dear and lovely Nathalie, I am grateful for the enormous support and patience you have shown me during the last couple of years.

Related publications

Journal papers

- Bentsen, T., A. A. Kressner, T. Dau, and T. May (2018). The impact of exploiting spectro-temporal context in computational speech segregation. *J. Acoust. Soc. Am.* 143.1, pp. 248-259. doi: 10.1121/1.5020273.
- Bentsen, T., T. May, A. A. Kressner, and T. Dau (2018). The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility. *PloS ONE* 13 (5): e0196924. https://doi.org/10.1371/journal.pone.0196924.
- Bentsen, T., S. Mauger, A. A. Kressner, T. May, and T. Dau (in review). The impact of noise power estimation on speech intelligibility in cochlear-implant speech coding strategies. *J. Acoust. Soc. Am.*, in review.

Conference papers

- Bentsen, T., T. May, A. A. Kressner, and T. Dau (2016). Comparing the influence of spectro-temporal integration in computational. speech segregation. *Proc. Interspeech*, 170-174, San Francisco, USA, 2016.
- May, T., T. Bentsen, and T. Dau (2015). The role of temporal resolution in modulation-based speech segregation. *Proc. Interspeech*, 170-174, Dresden, Germany, 2015.
- Kressner, A. A., T. May, R. M. Thaarup Høegh, A. K. Juhl, T. Bentsen, and T. Dau. Investigating the effects of noise-estimation errors in simulated cochlear implant speech intelligibility *Proc. ISAAR*, 6, 295-302, 2019.

Contents

1

Overall introduction

"When people talk, listen completely. Most people never listen."
—Ernest Hemingway

Communication is one of the most essential human skills. Being able to communicate requires listening to and understanding speech from other people. However, understanding speech in noise can be a challenge for many. In particular, hearing-impaired people demonstrate a poor speech understanding in the presence of competing talkers, since they are not able to listen in the valleys of fluctuating noise (Festen and Plomp, 1990). Even normal-hearing people can be challenged in conditions where the speech is corrupted by interfering noise at low signal-to-noise ratios (SNRs).

To address this challenge, modern communication devices such as hearing aids or cochlear-implants (CIs) make use of strategies to improve the speech understanding in noise. Devices with only one microphone rely on single-channel noise reduction. Devices with more than one microphone typically use multi-channel adaptive beamforming, based on an array of directional microphones, followed by post-filtering for a noise reduction (Zelinski, 1988; Simmer et al., 2001; Gannot and Cohen, 2008; Jensen and Pedersen, 2015). The performance of these single-channel strategies is therefore important to assess. Single-channel noise reduction strategies typically fail to improve speech intelligibility in adverse conditions, where the speech has been corrupted by competing talkers at a low SNR, for hearing-aid users (Dillon, 2001; Loizou et al., 2005; Hu and Loizou, 2007; Bentler et al., 2008; Loizou and Kim, 2011) or for CI recipients (Dawson et al., 2011; Mauger et al., 2012a). A goal is therefore to investigate and develop novel single-channel strategies which can improve speech intelligibility in adverse conditions. Besides being used in existing communication devices, these strategies can be employed in "hearables" that normal-hearing people can wear in noisy environments. Finally, the strategies

can be considered as a front-end in speech and speaker recognition systems to increase the robustness of such systems in noisy environments (Cooke et al., 2001; May et al., 2012a,b). These recognition systems are relevant in intelligent personal assistants.

The performance of the noise-reduction strategies depends on how well the characteristics of the speech and the noise signals are known, i.e. accurate estimates of the speech and the noise signals are needed. Therefore, it is necessary to have automatic approaches that can separate the speech from the noise as accurate as possible. However, separating speech from noise based on only a single channel input is a difficult task. For decades, researchers have attempted to solve this task using different approaches. One approach is to estimate the power of the noise signal based on statistical modeling of the underlying noise distribution (Martin, 2001; Hendriks et al., 2010; Gerkmann and Hendriks, 2012). The estimated noise power is then used to compute gain values which are applied to the noisy speech (Ephraim and Malah, 1984; Ephraim and Malah, 1985). These noise power estimators do not require pre-training for any specific acoustical condition, i.e. they are generic, and most of them are real-time applicable because of low latency values. Hence, they are suitable for practical applications in hearing aids or CIs. However, since the estimators are based on statistical assumptions about the underlying noise distribution, the accuracy of the noise power estimation can be a limiting factor in many conditions.

To overcome this limitation, Wang (2005) first presented the concept of an ideal time-frequency mask in which the level of speech activity is indicated in each time-frequency unit. The time-frequency mask is subsequently applied to a time-frequency representation of the noisy speech signal and converted back to the temporal waveform, via a synthesis step. Either an ideal binary mask (IBM) or an ideal ratio mask (IRM) can be constructed. In the IBM, the mask values are binary where time-frequency units with SNR exceeding an local criterion (LC) are considered speech-dominated and labeled one, whereas time-frequency units with SNR below the LC are considered to be noise-dominated and are labeled zero (Wang, 2005). The IRM mask values consist of continuous gain values between zero and one (Srinivasan et al., 2006; Narayanan and Wang, 2013; Hummersone et al., 2014; Wang et al., 2014).

In general, the ideal time-frequency mask is based on *a priori* information about the speech and the noise signals, and therefore needs to be estimated. This allows machine learning to be used in computational speech segregation where the ideal time-frequency mask can be considered the learning objective. Typically, a speech segregation system combines acoustic feature extraction with machine-learning techniques. The feature extraction is often based on principles of auditory processing in different frequency channels, assuming that auditory-inspired features are able to capture fundamental properties of the signals. Supervised learning is typically considered to achieve artificial intelligence, i.e. the machine-learning techniques are employed to make the speech segregation system capable of separating speech from noise without being explicitly instructed to do so. The distributions of the speech and the noise are learned through an initial training session. For the speech segregation systems to be applicable in practice, it is important to consider how accurately various acoustic conditions, which are not seen during the training session, can be predicted. This generalization ability is a focus point in computational speech segregation and a challenge for many systems.

To evaluate the single-channel noise-reduction strategies, objective measures have often been considered. Some of these objective measures compute the estimation accuracy by comparing to a priori knowledge (Kim et al., 2009; Hendriks et al., 2010; Gerkmann and Hendriks, 2012). Others predict the speech intelligibility by using models (Taal et al., 2011; Jensen and Taal, 2016). Measuring speech intelligibility in listeners is, however, important to properly assess approaches within computational speech segregation in the context of single-channel noise reduction.

This thesis investigates three approaches within computational speech segregation based on ideal time-frequency mask estimation. The approaches are all evaluated in the framework of single-channel noise reduction in normal-hearing listeners and CI recipients. The thesis is structured in three main chapters.

In *Chapter 2*, a speech segregation system is considered in which machine-learning techniques are employed for each auditory frequency channel (called a subband) to estimate the IBM. In speech, sentences are composed of words

that consist of syllables. These syllables may occupy several neighboring time-frequency units in the mask. Speech-dominated units are therefore clustered into these *glimpses* of speech in the mask, and the size of the glimpses correlates with speech intelligibility (Cooke, 2006; Barker and Cooke, 2007). The study focus is on how to exploit this contextual information in speech across time and frequency in computational speech segregation. Specifically, the impact of exploiting spectro-temporal context, through different strategies, is investigated on measured speech intelligibility in normal-hearing listeners. Furthermore, the generalization ability of the subband-based system is assessed by considering unseen noise segments in the evaluation. Measured speech intelligibility is finally compared to objective measures for the selection of a cost function that correlates with speech intelligibility. Supplementary work can be found in *Appendix A*. In Appendix A, the effect of changing the duration of the noise in the training session is investigated on the same range of objective measures.

In *Chapter 3*, the roles and the relative contributions of a selection of components within recent approaches in computational speech segregation are explored. Instead of subband-based approaches, similar to the one considered in Chapter 2, deep neural networks (DNNs) have been employed in the recent approaches and speech intelligibility improvements have been demonstrated in various adverse conditions (Healy et al., 2015; Chen et al., 2016a; Healy et al., 2017; Kolbæk et al., 2017). The considered components, which may be responsible for this success, are the system architecture, a time frame concatenation technique and the learning objective in computational speech segregation (i.e., the IBM or the IRM). The components are systematically investigated by measuring speech intelligibility in normal-hearing listeners. First, the architecture of a DNN-based system is compared to the architecture of the subband-based system presented in Chapter 2. Secondly, to exploit temporal context in the DNN-based system, the time frame concatenation technique is employed. Finally, the effect of IRM estimation versus IBM estimation is studied in the context of an otherwise identical DNN-based system.

In *Chapter 4*, an application of the estimated ideal time-frequency mask is considered in speech-coding strategies in real-time CI processing. Specifically,

a CI frequency channel is selected for electrical stimulation based on a SNR criterion. A frequency channel with a high instantaneous SNR conveys more reliable speech information than a frequency channel with a low instantaneous SNR, and only channels with SNRs larger or equal to an LC are therefore selected for stimulation. This is similar to how the IBM is constructed. In real-time CI processing, a noise power estimation stage is, however, required to estimate the instantaneous SNRs. Chapter 4 therefore investigated the impact of a state-of-the-art noise power estimator from Gerkmann and Hendriks (2012) in such speech-coding strategies. In addition, the role of the LC is investigated in the strategy, and compared to a speech-coding strategy using a fixed SNR-based channel selection. For the evaluation, speech intelligibility is measured and sound quality is rated in CI recipients in noisy conditions.

Finally, in *Chapter 5* the main findings are summarized and implications of the main findings are then discussed. In addition, perspectives for future studies are provided.

2

The impact of spectro-temporal context in computational speech segregation[a]

Abstract

Computational speech segregation aims to automatically segregate speech from interfering noise, often by employing ideal binary mask estimation. Several studies have tried to exploit contextual information in speech to improve mask estimation accuracy, by using two frequently-used strategies that (1) incorporate delta features and (2) employ support vector machine (SVM) based integration. In this study, two experiments were conducted. In Experiment I, the impact of exploiting spectro-temporal context using these strategies was investigated in stationary and six-talker noise. In Experiment II, the delta features were explored in detail, and tested in a setup that considered novel noise segments of the six-talker noise. Computing delta features led to higher intelligibility than employing SVM based integration and intelligibility increased with the amount of spectral information exploited via the delta features. A limited generalization ability was, however, observed with the computational speech segregation system. Measured intelligibility was subsequently compared to extended short-term objective intelligibility (ESTOI), hit - false alarm (H-FA) rate and the amount of mask clustering. None of these objective measures alone could account for measured intelligibility. The findings may have implications for the design of speech segregation systems, and for the selection of a cost function that correlates with intelligibility.

[a]This chapter is based on: Bentsen, T., A. A. Kressner, T. Dau, and T. May (2018). The impact of exploiting spectro-temporal context in computational speech segregation. *J. Acoust. Soc. Am.* 143.1, pp. 248-259. doi: 10.1121/1.5020273.

2.1 Introduction

The overall goal of computational speech segregation systems is to automatically segregate a target speech signal from interfering noise. These systems are relevant for many practical applications, e.g. as pre-processors in communication devices such as hearing aids or cochlear implants (Brungart et al., 2006; Li and Loizou, 2008; Wang et al., 2008) or front-ends in speech and speaker recognition systems for human-computer interfaces (Cooke et al., 2001; May et al., 2012a,b). One frequently-used single-channel approach, termed the ideal binary mask (IBM) technique (Wang, 2005), separates a time-frequency (T-F) representation of noisy speech into target-dominated and interference-dominated T-F units. Given *a priori* knowledge about the target and the interfering signal, the IBM is constructed by comparing the signal-to-noise ratio (SNR) in individual T-F units to a local criterion (LC). The resulting IBM is a binary matrix where T-F units with SNR exceeding the LC are considered target-dominated and labeled one, and zero otherwise. Many studies have used IBMs to segregate a target speech signal from a noisy mixture and demonstrated large intelligibility improvements (Brungart et al., 2006; Wang et al., 2008; Kjems et al., 2009). However, a priori knowledge about the target and the interfering noise is rarely available in realistic conditions, and therefore, the goal of segregation systems is to estimate the IBM based on the noisy speech signal. This challenge of obtaining an estimated IBM is typically approached by employing supervised learning strategies (Wang, 2005), which generally consist of a feature extraction front-end and a classification back-end. The front-end extracts a set of acoustic features which attempt to capture speech- and interference-specific properties. The distributions of speech and interference-dominated T-F units are then learned by a classification back-end, through an initial training stage (Kim et al., 2009; Han and Wang, 2012; Healy et al., 2013; May and Dau, 2014a).

When analyzing binary mask patterns, speech-dominated T-F units tend to cluster in spectro-temporal regions, forming so-called *glimpses*, and the size of these glimpses, denoted the glimpse proportion in the model by Cooke (2006), has been shown to correlate with speech intelligibility scores from normal-hearing listeners (Cooke, 2006; Barker and Cooke, 2007). Consequently, several studies have tried to exploit spectro-temporal contextual information

in speech to improve the performance of computational speech segregation systems by predominantly using two strategies. One strategy is to exploit the context in the front-end by calculating so-called *delta features* (Kim et al., 2009; Hu and Loizou, 2010; May and Dau, 2014b), which capture feature variations across time and frequency. Alternatively, the context can be exploited in the back-end where the posterior probability of speech presence obtained from a first classifier can be learned by a second classifier across a spectro-temporal window of T-F units, where the amount of spectro-temporal context can be controlled by the size of the window function (Han and Wang, 2012; Healy et al., 2013; May and Dau, 2014a). Some studies have combined both strategies in the front-end and in the back-end (Healy et al., 2013; May and Dau, 2013).

The performance of computational speech segregation systems and the effectiveness of different system configurations have primarily been evaluated based on the H - FA rate, which calculates the difference between the percentage of correctly classified speech-dominated T-F units (hit rate, H) and the percentage of incorrectly classified noise-dominated T-F units (false alarm rate, FA) (Kim et al., 2009; Han and Wang, 2012; Healy et al., 2013; May and Dau, 2013, 2014a; May and Dau, 2014b). However, it has recently been shown that speech intelligibility scores strongly depend on both the *distribution* of mask errors and the H - FA rate (Kressner and Rozell, 2015; Kressner and Rozell, 2016; Kressner et al., 2016). Specifically, Kressner and Rozell (2015) developed a graphical model to systematically measure the influence of clustering of T-F units on the intelligibility of binary-masked speech and showed that the intelligibility was reduced when masks contained an increased amount of clustering among T-F units, but the same mask error rates. Thus, the applicability of the H - FA rate as the sole objective measure to optimize or evaluate computational segregation systems has come into question. However, the impact of the different spectro-temporal context-exploring strategies on the amount of clustering of T-F units, or on speech intelligibility, has not yet been analyzed.

Kim et al. (2009) were the first to report speech intelligibility improvements for a computational speech segregation system based on Gaussian mixture models (GMMs). They considered a high complexity GMM classifier with 256 components in the back-end for modeling the distribution of the feature

vectors in a restricted setup in which the same short noise recording was used during training and testing. By using such a setup, it was possible to achieve high H - FA rates and improve speech intelligibility scores by up to 60% compared to unprocessed noisy speech for normal-hearing (NH) subjects (Kim et al., 2009). A high complexity classifier is able to learn all spectro-temporal characteristics of the noise, if the same short noise recording is used during training and testing, resulting in high H - FA rates (May and Dau, 2014b) and, most likely, also the high intelligibility scores observed in Kim et al. (2009). The restricted setup therefore has a high potential to improve speech intelligibility and can be used to investigate the behavior of the segregation system by comparing different system configurations. The ability of segregation systems to generalize to unseen acoustic conditions, such as novel segments of the same noise and novel noise types, is, however, an important and active research field (Healy et al., 2015; Chen et al., 2016a) and needs to be addressed at the same time.

In the present study, two experiments were conducted by measuring word recognition scores (WRSs) in NH listeners. In Experiment I, the impact of exploiting spectro-temporal context in the front-end and the back-end of a segregation system, based on GMMs, was systematically investigated to identify the best performing strategy for the system. Specifically, the extraction of the delta features (Kim et al., 2009) was considered in the front-end, and the two-layer classification stage from May and Dau (2014a) was employed in the back-end. Different system configurations were compared here, which either incorporated spectro-temporal context only in the front-end, only in the back-end or in both. These configurations were compared to a baseline configuration that did not include any of the strategies in the front-end and the back-end. This experiment was conducted in a restricted setup, similar to Kim et al. (2009), with high potential to improve speech intelligibility. Furthermore, the effect of the GMM classifier complexity in a segregation system was also investigated by comparing the results obtained with 16 GMM components and 64 GMM components. In Experiment II, the best performing strategy from Experiment I was explored in detail, and the generalization ability was subsequently evaluated in a less restricted setup that considered a mismatch in noise segments during training and testing. Finally, the intelligibility scores

from both experiments were related to predictions from objective measures[1] from the extended short-term objective intelligibility (ESTOI) (Jensen and Taal, 2016), the H - FA rate (Kim et al., 2009) and the amount of clustering among T-F units in binary masks (Kressner and Rozell, 2015). The primary focus of the later analysis was to guide the selection of a cost-function, that correlates with speech intelligibility, for future applications in computational speech segregation systems.

2.2 The segregation system

The segregation system consisted of a feature extraction front-end and a classification back-end (May et al., 2015). Figure 2.1 illustrates the processing stages of the system. Each of these stages is described in more detail below.

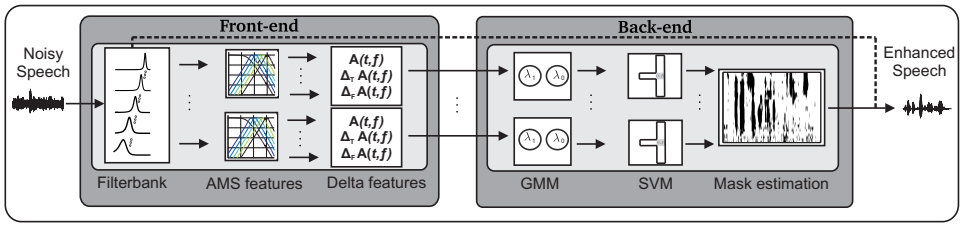


Figure 2.1: Block diagram of the speech segregation system. The system consists of a feature extraction front-end and a classification back-end. In the front-end, the noisy speech is first decomposed by a gammatone filterbank. Then, amplitude modulation spectrogram (AMS) features are extracted and delta features are computed. The back-end consists of two layers with a Gaussian mixture model (GMM) classifier in the first layer and a support vector machine (SVM) classifier in the second layer. Finally, the estimated ideal binary mask is applied to the subband signals of the noisy speech, as illustrated by the dashed line, in order to reconstruct the target signal.

2.2.1 Front-end

The noisy speech was sampled at a rate of 16 kHz and decomposed into $K = 31$ frequency channels by employing an all-pole version of the gammatone filterbank (Lyon, 1997), whose center frequencies were equally spaced on the equivalent rectangular bandwidth (ERB) scale between 80 and 7642 Hz. Previous

[1] Predictions in Experiment I were based on simulations with the objective measures and these predictions have been presented at the 17th Annual Conference of the International Speech Communication Association, San Francisco, USA and published as part of the conference proceedings in Bentsen et al. (2016).

studies (Kim et al., 2009; May and Dau, 2014a; May et al., 2015) have successfully exploited modulations in the speech and the interferer by extracting amplitude modulation spectrogram (AMS) features (Kollmeier and Koch, 1994; Tchorz and Kollmeier, 2003). To derive the AMS features in each subband, the envelope was extracted by half-wave rectification and low-pass filtering with a cutoff frequency of 1 kHz. Then, each envelope was normalized by its median computed over the entire envelope signal. The normalized envelopes were then processed by a modulation filterbank that consisted of one first-order low-pass and five band-pass filters with logarithmically spaced center frequencies and a constant Q-factor of 1. The cutoff frequency of the modulation low-pass filter was calculated as the inverse of the window duration to ensure that at least one full period of the modulation frequency was included in the window, and subsequently adjusted to the nearest power of 2 integer (May et al., 2015). Using a time frame duration of 32 ms then resulted in a cutoff frequency of 32 Hz. The root mean square (RMS) value of each modulation filter was then calculated across each time frame with a 75 % overlap. The extraction of the AMS features resulted in a six-dimensional feature vector for each T-F unit $\mathbf{A}(t,f) = \{M_1(t,f),\ldots,M_6(t,f)\}^T$. The delta features across time (Δ_T) and frequency (Δ_F) can be appended to the feature vector $\mathbf{A}(t,f)$ according to previous studies (Kim et al., 2009; Han and Wang, 2012; May and Dau, 2013), resulting in a feature vector $\mathbf{X}(t,f)$ for each individual T-F unit at time frame t and subband f that consists of:

$$\begin{aligned}
\mathbf{X}(t,f) &= \left[\mathbf{A}(t,f), \Delta_T\mathbf{A}(t,f), \Delta_F\mathbf{A}(t,f)\right] \\
\Delta_T\mathbf{A}(t,f) &= \begin{cases} \mathbf{A}(2,f) - \mathbf{A}(1,f), & \text{if } t = 1 \\ \mathbf{A}(t,f) - \mathbf{A}(t-1,f), & \text{otherwise} \end{cases} \\
\Delta_F\mathbf{A}(t,f) &= \begin{cases} \mathbf{A}(t,2) - \mathbf{A}(t,1), & \text{if } f = 1 \\ \mathbf{A}(t,f) - \mathbf{A}(t,f-1), & \text{otherwise} \end{cases}
\end{aligned} \tag{2.1}$$

Instead of the calculation in Eq. (2.1), delta features that only operate across frequency can be considered and appended symmetrically to the AMS features

for a resulting feature vector $\mathbf{X}(t,f)$:

$$\begin{aligned}\mathbf{X}(t,f) &= \left[\mathbf{A}(t,f), \Delta_{f-k}\mathbf{A}(t,f), \Delta_{f+k}\mathbf{A}(t,f)\right] \\ \Delta_{f-k}\mathbf{A}(t,f) &= \mathbf{A}(t,f) - \mathbf{A}(t,f-k), \quad \forall k \in \{n \in [1;K] | f-n \geq 1\} \\ \Delta_{f+k}\mathbf{A}(t,f) &= \mathbf{A}(t,f) - \mathbf{A}(t,f+k), \quad \forall k \in \{n \in [1;K] | f+n \leq K\}\end{aligned} \tag{2.2}$$

In Eq. (2.2), k indicates the considered number of subbands in the calculation, and K the number of gammatone filters. Appending the delta features to the feature vector in Eqs. (2.1) and (2.2) increased the amount of exploited spectro-temporal context, but also the size of the feature vector. E.g., appending $\Delta_T\mathbf{A}(t,f)$ and $\Delta_F\mathbf{A}(t,f)$ to $\mathbf{A}(t,f)$ in Eq. (2.1) would increase the feature vector from 6 to 18 dimensions.

2.2.2 Back-end

Similar to previous studies, the classification back-end consisted of a two-layer segregation stage (Healy et al., 2013; May and Dau, 2014a; May et al., 2015). In the first layer, a GMM classifier was trained to represent the speech- and noise-dominated AMS feature distributions ($\lambda_{1,f}$ and $\lambda_{0,f}$) for each subband f. To separate the feature vector into speech- and noise-dominated T-F units, the LC was applied to the *a priori* SNR, and the *a priori* probabilities $P(\lambda_{1,f})$ and $P(\lambda_{0,f})$ were computed by counting the number of feature vectors for each of the classes $\lambda_{1,f}$ and $\lambda_{0,f}$ during training. The GMM classifier output was given as the posterior probability of speech and noise presence $P(\lambda_{1,f}|\mathbf{X}(t,f))$ and $P(\lambda_{0,f}|\mathbf{X}(t,f))$, respectively:

$$P(\lambda_{1,f}|\mathbf{X}(t,f)) = \frac{P(\lambda_{1,f})P(\mathbf{X}(t,f)|\lambda_{1,f})}{P(\mathbf{X}(t,f))} \tag{2.3}$$

$$P(\lambda_{0,f}|\mathbf{X}(t,f)) = \frac{P(\lambda_{0,f})P(\mathbf{X}(t,f)|\lambda_{0,f})}{P(\mathbf{X}(t,f))} \tag{2.4}$$

For each subband, the computed posterior probabilities of speech $P(\lambda_{1,f}|\mathbf{X}(t,f))$ were processed by a linear support vector machine (SVM) classifier (Chang and Lin, 2011) across a spectro-temporal window $\mathcal{W}$ (May and Dau, 2014a):

$$\tilde{\mathbf{X}}(t,f) = \left\{P(\lambda_{1,u}|\mathbf{X}(u,v)) : (u,v) \in \mathcal{W}(t,f)\right\}. \tag{2.5}$$

The size of the window $\mathcal{W}$ determined the amount of spectro-temporal context exploited around the considered T-F unit. A causal and plus-shaped window function $\mathcal{W}$ was used here, where the window size with respect to time and frequency was controlled by Δt and Δf, respectively. Further details regarding the choice of the second-layer classifier and the size and shape of the window function $\mathcal{W}$ can be found in May and Dau (2014a).

2.3 Methods

2.3.1 Configurations

To systemically analyze the impact of spectro-temporal context strategies in the front-end and the back-end, four system configurations were tested in Experiment I, see Table 2.1. The "No context" configuration denotes the baseline configuration with no delta features in the front-end and no spectro-temporal integration in the back-end, corresponding to setting the window size $\mathcal{W}$ to unity ($\Delta t = 1, \Delta f = 1$). The "Front-end" configuration includes the delta features, while the "Back-end" configuration includes the second-layer classification stage in the back-end ($\Delta t = 3, \Delta f = 9$). The "Front- & back-end" configuration employs both the front-end and the back-end spectro-temporal context strategies.

Table 2.1: Configurations in Experiment I

Configurations	Front-end		Back-end	
	Feature vector $\mathbf{X}(t,f) =$	Feature dimension	$\mathcal{W}$ size Δt	$\mathcal{W}$ size Δf
No context	$[\mathbf{A}(t,f)]$	6	1	1
Front-end	$[\mathbf{A}(t,f), \Delta_T\mathbf{A}(t,f), \Delta_F\mathbf{A}(t,f)]$	18	1	1
Back-end	$[\mathbf{A}(t,f)]$	6	3	9
Front- & back-end	$[\mathbf{A}(t,f), \Delta_T\mathbf{A}(t,f), \Delta_F\mathbf{A}(t,f)]$	18	3	9

In Experiment II, the delta features were explored in detail in order to investigate the potential of this strategy in the segregation system. Four configurations were selected, see Table 2.2. The system configuration "Front-end" is the baseline configuration for the analysis across frequency and appends only $\Delta_F\mathbf{A}(t,f)$ to $\mathbf{A}(t,f)$. The configurations "3 subbands", "7 subbands" and "11 subbands" append $k = 1$, $k = 3$ and $k = 5$ lower and upper subbands to $\mathbf{A}(t,f)$.

Table 2.2: Configurations in Experiment II

Configurations	Front-end	
	Feature vector $\mathbf{X}(t,f) =$	Feature dimension
Front-end	$[\mathbf{A}(t,f), \Delta_F \mathbf{A}(t,f)]$	12
3 subbands	$[\mathbf{A}(t,f), \Delta_{F-1}\mathbf{A}(t,f), \Delta_{F+1}\mathbf{A}(t,f)]$	18
7 subbands	$[\mathbf{A}(t,f), \Delta_{F-1}\mathbf{A}(t,f), \Delta_{F+1}\mathbf{A}(t,f), ..., \Delta_{F+3}\mathbf{A}(t,f)]$	42
11 subbands	$[\mathbf{A}(t,f), \Delta_{F-1}\mathbf{A}(t,f), \Delta_{F+1}\mathbf{A}(t,f), ..., \Delta_{F+5}\mathbf{A}(t,f)]$	66

2.3.2 Stimuli

The speech material came from the Danish Conversational Language Understanding Evaluation (CLUE) database (Nielsen and Dau, 2009). It consists of 70 sentences in 7 lists for training and 180 sentences in 18 balanced lists for testing, and is spoken by a male Danish talker. Noisy speech mixtures were created by mixing individual sentences with a stationary (ICRA1) and a fluctuating six-talker (ICRA7) noise (Dreschler et al., 2001). A Long Term Average Spectrum (LTAS) template was computed based on the CLUE corpus and the LTAS of each noise masker was adjusted to the template LTAS. A randomly-selected noise segment was used for each sentence. In order to avoid onset effects in the speech intelligibility test (Nielsen and Dau, 2009), the noise segment started 1000 ms before the speech onset and ended 600 ms after the speech offset. However, the objective measures were computed only for the regions between speech onset and offset.

2.3.3 System training and evaluation

In Experiment I, the segregation system was trained separately for the two noise types limited to 10 s in duration. Originally, the ICRA1 consists of a 60 s noise recording and ICRA7 of a 600 s recording (Dreschler et al., 2001). The first layer of the classification back-end consisted of a subband GMM classifier with either 16 or 64 components and full covariance matrices. The classifiers were first initialized by 15 iterations of the K-means clustering algorithm, followed by 5 (for 16 GMMs) or 50 (for 64 GMMs) iterations of the expectation-maximization algorithm. The classifiers were trained with the 70 training sentences that were each mixed three times with a randomly-selected noise segment from 10 s noise recordings at $-5, 0$, and 5 dB SNR. The subsequent linear SVM classifier was trained for each subband with only 10 sentences mixed at $-5, 0$, and 5 dB SNR.

Afterwards, a re-thresholding procedure was applied (Han and Wang, 2012; May and Dau, 2014a) using a validation set of 10 sentences, where new SVM decision thresholds were obtained which maximized the H - FA rates. Both the first and second-layer classifiers employed a LC of −5 dB in a similar manner as previous findings (Han and Wang, 2012; May and Dau, 2014b). The segregation system was evaluated with the 180 CLUE sentences. Each sentence was mixed with the noises at −5 dB SNR using the same limited noise recordings from the training session.

Experiment II only tested the highly non-stationary ICRA7 noise type in a less restricted setup. This noise type is more likely to challenge a speech segregation system than the stationary ICRA1. The full noise recording of 600 s was divided into one half recording for training and one half recording for testing. The training and evaluation was similar to Experiment I. The first layer of the classification back-end had a complexity of 16 Gaussian components with full covariance matrix. The complexity choice is discussed in Sec. 2.5.2.

2.3.4 Test procedure and subjects

In Experiment I, the following 24 conditions were tested: (Noisy speech, No integration, Front-end, Back-end, Front- & back-end, IBM) X (ICRA1, ICRA7) X (16 GMM components, 64 GMM components). The total number of conditions (24) exceeded the number of available CLUE test lists (18). Therefore, to be able to randomly assign one condition to one test list, the experiment was conducted with two subject groups, each with $n = 15$ NH listeners. The first subject group was tested with the 12 conditions corresponding to the classifier complexity of 16 GMMs, and the second group was tested with the 12 conditions with only 64 GMMs. The following 5 conditions were tested in Experiment II: Noisy speech, Front-end, 3 subbands, 7 subbands & 11 subbands. The experiment was conducted with one subject group with $n = 20$ NH listeners that differed from the subject groups used in Experiment I. In this experiment, 13 other conditions were also tested that were not relevant to this study.

The listener age was between 20 and 32 years with a mean of 24.5 years in Experiment I and a mean of 26.7 years in Experiment II. Requirements for participation were: (1) age between 18 − 40 years, (2) audiometric thresholds of less than or equal to 20 dB hearing level (HL) in both ears (0.125 to 8 kHz), (3)

Danish as native language, and (4) no previous experience with the Hearing In Noise Test (HINT) (Nielsen and Dau, 2011) or CLUE (Nielsen and Dau, 2009). The total experimental time was about 2 hours in Experiment I and about 1.5 hour in Experiment II, including the screening process. The experiments were approved by the Danish Science-Ethics Committee (reference H-16036391), and the subjects were paid for their participation.

The experiments consisted of a training and testing session. During the training session, 5 randomly selected sentences from the training set were presented for each of the 12 conditions to familiarize the subject to the task. Subsequently, each subject heard one list per condition, and conditions and lists were randomized across subjects. The sentences were presented diotically to the listener via headphones (Sennheiser HD650) in an acoustically and electrically shielded booth. Prior to the actual experiments, the headphones were calibrated by first adjusting to a reference sound pressure level (SPL) value and then performing a headphone frequency response equalization. During the experiment, the sentences were adjusted to the desired presentation level, and the equalization filters were applied. The SPL was set to a comfortable level of 65 dB. The presentation level was only increased after the training session if the subject reported back that the level was too low. The level never exceeded 70 dB SPL for any subject. For each sentence, the subjects were instructed to repeat the words they heard, and an operator scored the correctly understood words via a MATLAB interface. The subjects were told that guessing was allowed. They could listen to each sentence only once, and breaks were allowed according to the subject's preference.

2.3.5 Statistical analysis

Intelligibility scores were reported as a percentage of correctly scored words, i.e. the WRSs, at -5 dB SNR. The WRSs were computed per sentence and averaged across sentences per list. The averaged WRSs were used to construct a linear mixed effect model for each experiment. In Experiment I, the three fixed factors of the mixed model were the system configuration (4 levels), the noise type (2 levels) and the classifier complexity (2 levels). The subjects were treated as a random factor, as is standard in a repeated measure design. The intelligibility scores in Experiment I followed a normal distribution. All fixed effects, all interactions between fixed effects and the random effect were initially

included in the model. The model was then reduced by performing a backward elimination of all random and fixed interactions that were non-significant. This included all of the interaction terms between the random effect (subjects) and the fixed factors (configuration, noise type and classifier complexity) and the interaction term between all three fixed factors. In Experiment II, the only fixed factor was system configuration (4 levels) and subjects were treated as a random factor. The intelligibility scores in Experiment II also followed a normal distribution.

All levels were tested at a 5% significance level. To visualize the data, the least-squares means and 95% confidence intervals were extracted from the model. To assess any difference between conditions, the differences of the least-squares means were computed and the p values were adjusted following the Tukey multiple comparison testing. To evaluate potential speech intelligibility improvements, Paired Student's t-tests between the noisy speech and each of the system configurations were constructed and tested at a 5% significance level.

2.3.6 Objective measures

Three different objective measures were compared to the intelligibility scores in each experiment: ESTOI (Jensen and Taal, 2016), H - FA rate (Kim et al., 2009) and the clustering parameter γ (Kressner and Rozell, 2015). The ESTOI (Jensen and Taal, 2016) is a modified version of the short-term objective intelligibility (STOI) index (Taal et al., 2011) to better account for modulated noise maskers. The STOI index is based on a short-term correlation analysis between the clean and the degraded speech (Taal et al., 2011), mapped to a value between 0 and 1. The ESTOI improvements (Δ ESTOI) were reported here as the relative difference between the predicted ESTOI values for the processed and the unprocessed noisy speech baselines. To compute the H - FA rate, the correctly classified speech-dominated T-F units and incorrectly classified noise-dominated T-F units were derived by comparing the estimated IBM with the IBM. The H - FA rates and the ESTOI improvements were averaged across all 180 test sentences. The clustering parameter γ was learned across all 180 test sentences by the graphical model described in Kressner and Rozell (2015). Given a set of binary masks, the graphical model estimates the amount of clustering γ between T-F units within the masks as a single number. γ quantifies how much more likely

neighboring T-F units are to have the same label (speech-dominated or noise-dominated) as opposed to different labels. Therefore, binary masks with T-F units that are twice as likely to have the same label than a different label as their neighboring units would be described by $\gamma = 2.0$. Binary masks with T-F units that are equally likely to be in the same state as their neighbors would have a $\gamma = 1.0$, indicating that the labels of the T-F units would be uniformly and randomly distributed. Therefore, a mask with $\gamma = 2.0$ will contain more clustering among the T-F units than a mask with $\gamma = 1.0$ (Kressner and Rozell, 2015). To illustrate the γ parameter, Fig. 2.2 shows binary masks for one particular CLUE sentence mixed with ICRA7 noise at -5 dB SNR with the respective γ values, shown in parenthesis. Figure 2.2a shows the IBM and Fig. 2.2b-e present the estimated IBMs for the four tested system configurations listed in Table 2.1. The two mask error types, misses and false alarms, are shown on top of the binary masks for a visualization of the error distributions. Comparing the masks for the four tested system configurations, the masks from Fig. 2.2d and Fig. 2.2e contain a larger amount of clustering than the masks in Fig. 2.2b and Fig. 2.2c.

2.4 Results

2.4.1 Experiment I: Impact of exploiting spectro-temporal context

Figure 2.3 shows intelligibility scores obtained with the four system configurations ("No Context", "Front-end","Back-end" and "Front- & back-end") in the two noise types (ICRA1 and ICRA7) considered in Experiment I. Results are shown for the two classifier complexities, namely 16 GMMs in Fig. 2.3a and 64 GMMs in Fig. 2.3b. The condition with the unprocessed noisy speech (diamonds) represented the baseline, and the IBM condition (stars) was considered as the ideal reference. For the baseline and the ideal reference, sample means across subjects and 95% Student's t-based confidence intervals of the mean were computed. For the system configurations, the least square means and 95% confidence intervals from the fitted linear mixed effect model were considered.

The baseline in Fig. 2.3 differed across noise types, with WRS of about 50−55% for the stationary ICRA1 and 65% for the fluctuating ICRA7, presumably because the participants were able to listen in the dips in the six-talker noise.

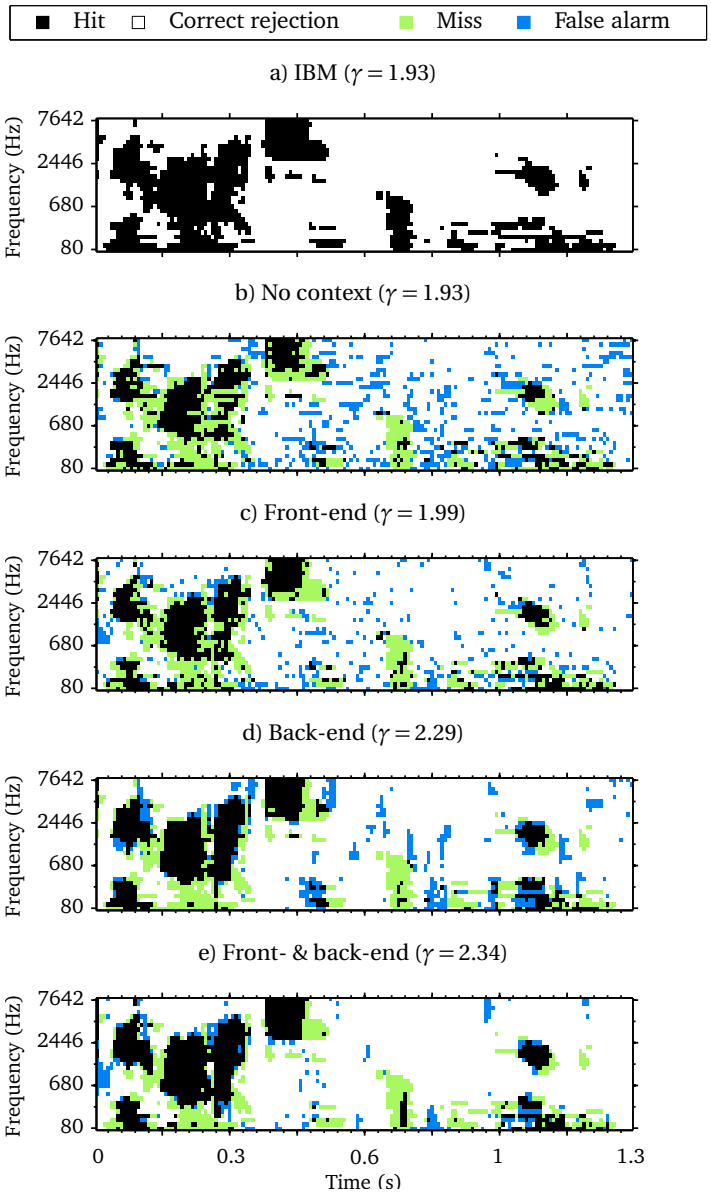


Figure 2.2: Binary masks for a CLUE sentence mixed with ICRA7 noise at −5 dB SNR. Misses (target-dominated T-F units erroneously labeled as masker-dominated) and false alarms (masker-dominated T-F units erroneously labeled as target-dominated) are shown on top of the masks. Similar panels are shown in Fig. A.3 in Bentsen et al. (2016) with another example sentence.

For the IBM conditions, WRS of close to 100% were achieved for both noise types. This was expected as the IBM exploited the *a priori* information about the speech and the noise signals.

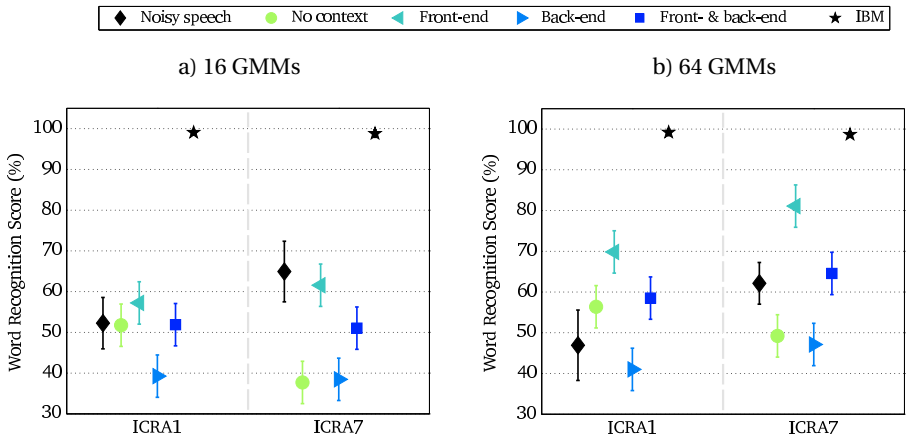


Figure 2.3: Experiment I's WRSs at −5 dB SNR of the four different system configurations ("No Context", "Front-end","Back-end" and "Front- & back-end") for the two noise types (ICRA1 and ICRA7) and for the two classifier complexities plotted in panel a) (16 GMMs) and panel b) (64 GMMs). The condition with the unprocessed noisy speech represented the baseline and the IBM condition was considered as the ideal reference. For the baseline and the ideal reference, sample means across subjects and 95% Student's t-based confidence intervals of the mean were computed. For all system configurations in all combinations of noise type and classifier complexity, the least square means and 95% confidence intervals from the fitted linear mixed effect model were plotted.

There was an effect of system configuration depending on the classifier complexity and on the noise type. Most importantly, the "Front-end" configuration led to significantly higher intelligibility scores than the "Back-end" configuration for both noise types and both classifier complexities ($p < 0.0001$). Specifically, the WRS increased by 18.0% in ICRA1 and 23.1% in ICRA7 with 16 GMMs (Fig. 2.3a), and 28.8% in ICRA1 and 34.0% in ICRA7 with 64 GMMs (Fig. 2.3b). This particular finding suggests that extracting and appending the delta features to the AMS features in the front-end is a more effective way of exploiting spectro-temporal contextual information than using the SVM-based integration strategy in the back-end. In all four combinations, except with 16 GMMs in the case of the ICRA1 noise, the "Front-end" configuration led to significantly larger scores than the "No context" configuration, which emphasizes that it is more effective to exploit contextual information in the front-end of the system than not considering any strategy at all. Finally, the "Front- & back-end" configuration also led to significantly higher scores than the "Back-end" configuration in all four combinations of noise type and classifier complexity. However, the mean scores for the "Front- & back-end" were generally lower than for the "Front-end". This suggests that employing

both strategies is more effective to exploit spectro-temporal context than just employing the SVM-based integration strategy in the back-end alone, but the combination of the two strategies does not lead to better results than the front-end strategy alone.

There was also an effect of the classifier complexity that depended on the system configuration and the noise type. By comparing the results in Fig. 2.3a and Fig. 2.3b, significantly higher scores were obtained for the "Front-end" configuration with 64 GMMs than with 16 GMMs for both noise types. Specifically, the WRS increased by 12.6% in ICRA1 ($p < 0.05$) and 19.5% in ICRA7 ($p < 0.0001$). However, the scores for the "Back-end" configuration did not change significantly across classifier complexity for either noise type. Most importantly, the ranking of the system configurations remained unchanged across classifier complexity.

The measured intelligibility scores from Fig. 2.3 were converted into WRS improvements relative to the unprocessed noisy speech, ΔWRS. Figure 2.4a and Fig. 2.4b show ΔWRS as a function of the system configuration, noise type and classifier complexity. Significant improvements, based on the Paired Student's t-tests, are indicated by an asterisk (*). Significant improvements of about 50% for ICRA1 and 35% for ICRA7 over noisy speech were obtained with the IBM. For 64 GMMs in Fig. 2.4b, the configurations "No Context" ($t[14] = -2.16, p = 0.02$), "Front-end" ($t[14] = -4.29, p =< 0.001$) and "Front- & back-end" ($t[14] = -2.82, p = 0.007$) for ICRA1 led to significant improvements and for the ICRA7, only the "Front-end" ($t[14] = -7.44, p =< 0.001$) led to a significant improvement. To evaluate the potential of the objective measures, the measured intelligibility scores were related to predictions from each of the objective measures described in Sec 2.3.6. Figure 2.4 also shows the objective measures ΔESTOI (Figs. 2.4c and 2.4d), H - FA rates (Figs. 2.4e and 2.4f) and γ (Figs. 2.4g and 2.4h) in Experiment I. ΔESTOI indicates the increase in ESTOI relative to the unprocessed noisy speech. The largest predicted improvement was observed for the configuration "Front- & back-end", and the lowest predicted improvement was found for the "No context" configuration in all combinations of noise type and classifier complexity level. This is in conflict with the measured ΔWRS in Figs. 2.4a and 2.4b where the "Front-end" configurations led to the largest improvements. By comparing Figs. 2.4c

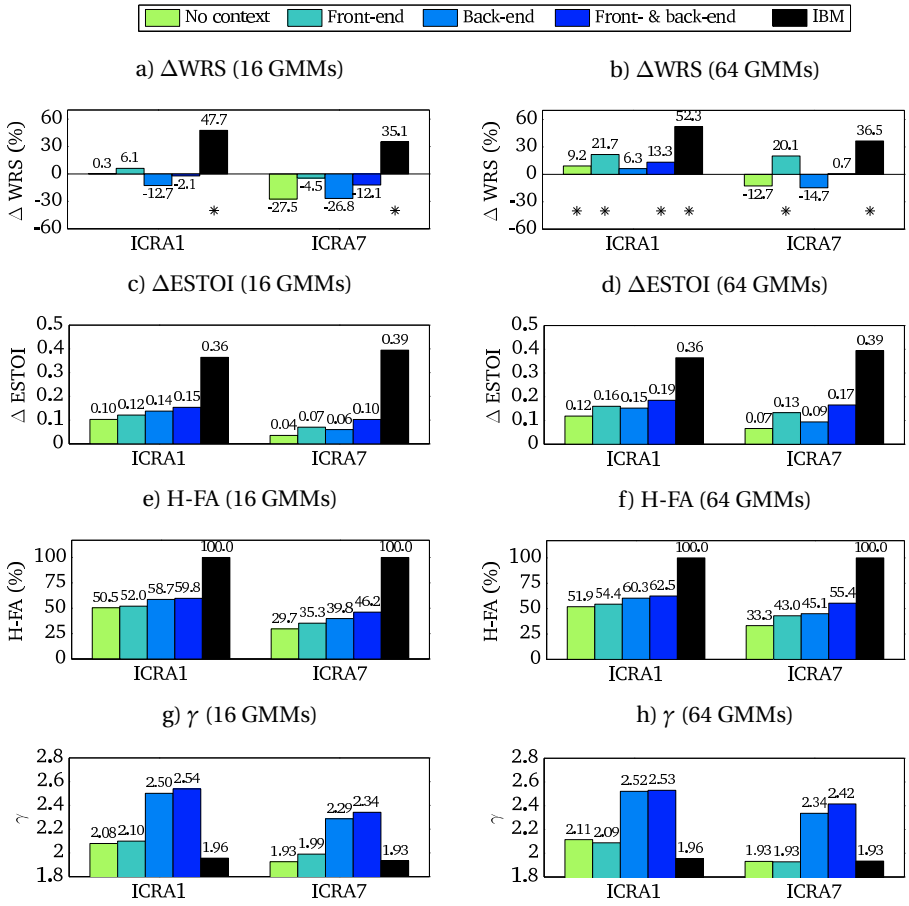


Figure 2.4: Experiment I's ΔWRS relative to noisy speech (first row of panels), ΔESTOI relative to noisy speech (second row of panels), H-FA rates (third row of panels) and γ values (fourth row of panels) for the four different system configurations with the two noise types (ICRA1 and ICRA7) and with the two classifier complexities in a) and in b). The IBM has been included as the ideal reference. WRS improvements are derived from the Paired Student's t-tests and significant improvements (on a 5% significance level) are marked with an asterisk (*). All objective measures are evaluated at −5 dB SNR.

and 2.4d, it can be seen that larger ESTOI improvements were generally observed with 64 GMMs compared to 16 GMMs. This is consistent with the measured WRS improvements in Figs. 2.4a and 2.4b.

Figures 2.4e and 2.4f show the H - FA rates. The segregation system generally produced higher H - FA rates in the presence of the stationary noise than in the presence of the non-stationary six-talker noise. The six-talker noise contains spectro-temporal modulations, similar to modulations in the target speech signal, and it will be more difficult for the classifier to separate the speech

modulations from the six-talker noise modulations. In all combinations of noise type and classifier complexity, the lowest H - FA rates were observed for the "No context" configuration and the highest H - FA rates were found for the "Front- & back-end" configuration. Also, larger H - FA rates were obtained for the "Back-end" than for the "Front-end" configuration, which is not consistent with Figs. 2.4a and 2.4b. Furthermore, higher H - FA rates were obtained with 64 GMMs in Fig. 2.4f than with 16 GMMs in Fig. 2.4e. A comparison with the measured WRS improvements in Figs. 2.4a and 2.4b indicated a conflict with this prediction, since the "Front-end" configuration led to the highest intelligibility scores, but not the highest H - FA rates. Finally, it is observed that a small increase of H - FA (from Fig. 2.4e to 2.4f) corresponds to a large increase of WRS (from Fig. 2.4a to 2.4b) from 16 GMMs classifier to the 64 GMMs classifier. This was found for both noise types.

Figures 2.4g and 2.4h show the γ values learned by the graphical model. The IBM itself contains a certain level of clustering, due to the compact representation of speech-dominated T-F units forming glimpses of the target signal. The γ values from system configurations that exploited spectro-temporal context through the SVM based integration strategy in the back-end ("Back-end" and "Front- & back-end") were consistently larger than the γ values learned over masks from the "Front-end" and the "No context" configurations. Furthermore, the "Front-end" did not lead to larger γ values than the "No context". This suggests that computing delta features in the front-end does not increase the amount of clustering, in contrast to employing a spectro-temporal SVM based integration strategy in the back-end. The effect of exploiting spectro-temporal context in binary masks was visualized in Fig. 2.2 in Sec. 2.3. Figure 2.2d-e showed masks with a larger amount of T-F clustering than the masks in Fig. 2.2b-c, and a visual inspection of the example utterance indicated that the erroneous T-F units became more clustered in Fig. 2.2d-e. Finally, a comparison of Fig. 2.4g and Fig. 2.4h suggests that the amount of clustering in the mask is not affected by the classifier complexity in the segregation system, as γ remains unchanged.

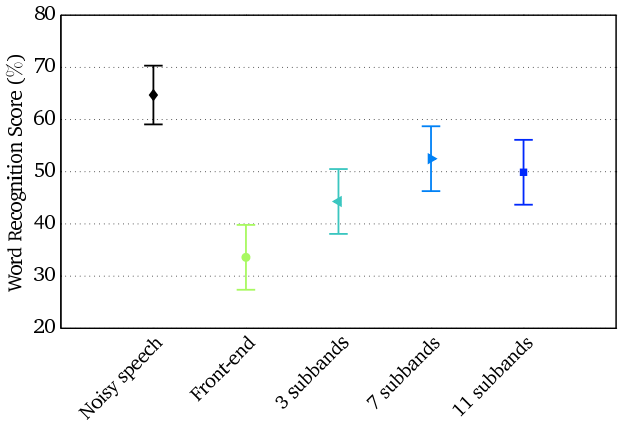


Figure 2.5: Experiment II's WRSs at −5 dB SNR with the four different system configurations ("Front-end", "3 subbands", "7 subbands" and "11 subbands") in ICRA7. The condition with the unprocessed noisy speech represented the baseline. For the baseline, sample means across subjects and 95% Student's t-based confidence intervals of the mean were computed. For all system configurations, the least square means and 95% confidence intervals from the fitted linear mixed effect model were plotted.

2.4.2 Experiment II: Exploring delta features and the system generalization ability

Figure 2.5 shows intelligibility scores obtained in Experiment II with the four system configurations ("Front-end", "3 subbands","7 subbands" and "11 subbands") tested in the less restricted setup in ICRA7 noise. For all four configurations, the $\Delta_T \mathbf{A}(t,f)$ from Eq. (2.1) was not appended to the feature vector in Eq. (2.2). This decision was based on an analysis of the objective measures prior to Experiment II, which showed no change in the objective measures when $\Delta_T \mathbf{A}(t,f)$ was left out. In Fig. 2.5, the level of the noisy speech was consistent with the level in Experiment I for ICRA7 (see Fig. 2.3). In this experiment, there was an effect of system configuration. The intelligibility scores were significantly higher in the "3 subbands" configuration than the "Front-end" configuration by 10.7% ($p < 0.01$) and from the "3 subbands" to the "7 subbands" configuration by 8.2% ($p < 0.05$). The "7 subbands" and the "11 subbands" configurations did not differ significantly. This finding indicated that appending more subbands, as proposed in Eq. (2.2), can lead to significantly higher intelligibility until a plateau at $k = 3$ with '7 subbands". Figure 2.6 presents the intelligibility improvements and objective measure predictions for Experiment II. In Fig. 2.6a, the Paired Student's t-tests showed

that all system configurations led to significantly smaller intelligibility scores than the noisy speech, despite an increase in intelligibility over appended subbands. Therefore, none of the system configurations were able to improve speech intelligibility in the less restricted setup. Since this setup included novel noise segments in testing not seen during training, this suggested that the segregation system had a limited ability to generalize to unseen noise segments of the six-talker noise.

In Fig. 2.6b, all predicted ΔESTOI values were positive, and the largest predicted improvements were observed for the configurations "7 subbands" and "11 subbands". This was not consistent with results from the listener study in Fig. 2.6a where no WRS improvements were observed, which highlights the discrepancy between predicted and measured intelligibility improvements in this study. The H - FA rate in Fig. 2.6c increased with the number of appended subbands, whereas the rates were comparable for "7 subbands" and "11 subbands". As observed in Experiment I, a small change in H - FA had a large impact on the measured intelligibility scores. This was illustrated by comparing Fig. 2.4e for the ICRA7 noise and Fig. 2.6c. A H - FA rate of 35.3% in Fig. 2.4e corresponded to a 4.5% decrease in WRS for the "Front-end" configuration whereas a H - FA of 33.6% in Fig. 2.6c corresponded to a 31.1% decrease in WRS over noisy speech. With respect to clustering (Fig. 2.6d), γ did not change with the system configuration, suggesting that the amount of clustering in the mask is not affected by appending more subbands to the AMS features. This is in contrast to the Experiment I where the SVM integration stage in the back-end increased both H - FA and γ.

2.5 Discussion

2.5.1 The impact of exploiting spectro-temporal context

The measured intelligibility scores in Experiment I (Sec. 2.4.1) showed that the front-end strategy, where the system was given access to both the AMS features and the delta features, led to significantly higher intelligibility scores than employing the back-end strategy, which incorporated the SVM-based spectro-temporal integration. The scores were consistently higher for the front-end strategy than the back-end strategy, regardless of the noise type and

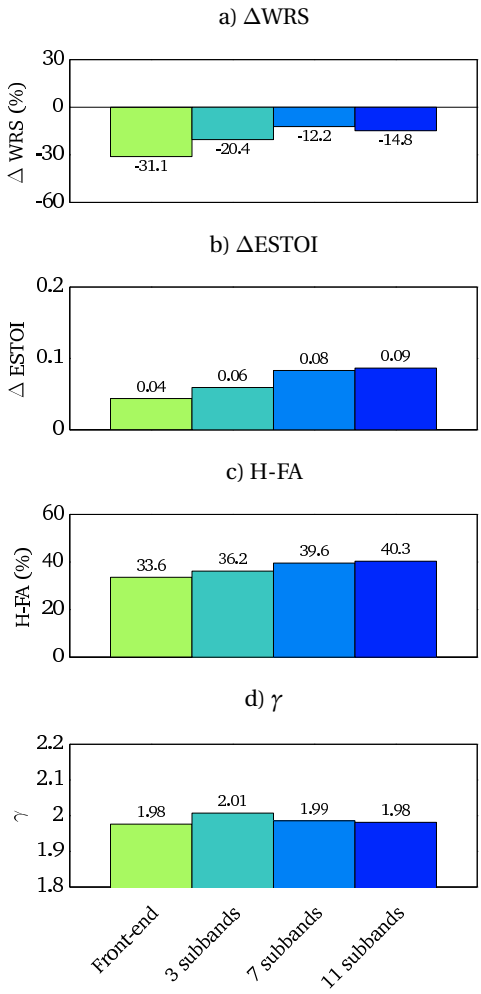


Figure 2.6: Experiment II's ΔWRS relative to noisy speech (first row of panels), ΔESTOI relative to noisy speech (second row of panels), H-FA rates (third row of panels) and γ values (fourth row of panels) with the four different system configurations in ICRA7. WRS improvements are derived from the Paired Student's t-tests and significant improvements (on a 5% significance level) are marked with an asterisk (*). All objective measures are evaluated at −5 dB SNR.

classifier complexity. Moreover, compared to the unprocessed noisy speech, the back-end strategy actually had a detrimental effect on the intelligibility scores. The comparison of the objective measures in Fig. 2.4 (Sec. 2.4.1) indicated that the back-end strategy increased the H - FA rates over the front-end strategy but, at the same time, increased the amount of clustering of individual T-F units. The visual inspection of the illustrated mask examples in Fig. 2.2 (Sec. 2.3.6) furthermore suggested that the increased amount of clustering implied an increased clustering of the misses and false alarms. Previously, it was shown

that clustering of the two error types results in reduced intelligibility scores despite having the same classification accuracy (Kressner and Rozell, 2015), which may explain the detrimental effect of the back-end strategy on the present intelligibility scores. Furthermore, computing delta features in the front-end had a positive effect on speech intelligibility. The intelligibility scores were significantly higher than the scores with the configuration that did not employ any of the strategies, and improvements over noisy speech were significant for the higher complexity classifier of 64 GMMs. Because of the detrimental effect of the back-end strategy on intelligibility, combining both strategies simultaneously in the front-end and in the back-end did not lead to the largest measured intelligibility scores in Sec. 2.4.1. This contradicted the findings in Fig. 2.4e and Fig. 2.4f (Sec. 2.4.1) where a higher H - FA rate was found when combining the strategies than employing only one of the strategies, consistent with the literature (Healy et al., 2013; May and Dau, 2013). The results from Experiment I therefore suggest that, in the considered segregation system, a better spectro-temporal strategy is to compute delta features of the AMS features in the front-end rather than employing the selected SVM-based integration strategy in the back-end. This study, however, did not consider the effects of changing the shape and the size of the window in the back-end on measured intelligibility. Also, the effect of employing a different second-layer classifier is currently unknown. Healy et al. (2013) considered a similar two-layer classification stage, but they employed deep neural networks (DNNs) in a DNN-DNN layer with an integration window of size 5 time frames and 17 subbands of the 64 channels. They reported significant improvements in intelligibility scores with this system, but did not quantify the impact of the back-end strategy alone.

In Experiment II, the front-end strategy was explored in detail by appending delta features computed from symmetrical subbands. Results in Sec. 2.4.2 showed that the intelligibility scores increased with the number of appended subbands up to $k = 3$ bands where the improvement reached a plateau. This indicated that intelligibility increased with the amount of spectral information in the speech that was exploited up to $k = 3$ subbands. The same trend was observed for the H - FA rate in Fig. 2.6. Appending the delta features across frequency increased the size of the feature vector, and the larger amount of training data led to improvements in H - FA rate for the higher complexity

classifier of 64 GMMs compared to the 16 GMMs classifier. Moreover, the amount of clustering among the T-F units in Experiment II was equal to the amount of clustering for the front-end strategy in Experiment I and remained constant with the number of appended subbands. This is in line with the notion from Experiment I that increased accuracy without increased clustering among the T-F units can lead to higher intelligibility scores.

Other strategies exists that exploit the contextual information in speech. In contrast to the delta features, which work on a subband level, temporal context can also be exploited by stacking feature frames as input to broadband DNNs for classification (Wang et al., 2014; Chen et al., 2016a). However, the impact of this particular strategy on intelligibility scores, or any of the objective measures, has not been quantified, which makes a comparison to the strategies in the present study challenging.

2.5.2 The generalization ability of the segregation system

In Experiment I, a restricted setup from Kim et al. (2009), with matched noises during training and testing, was used in order to facilitate a comparison of the system configurations, and for a comparison across GMM classifier complexity. May and Dau (2014b) compared H - FA rates for matched and mismatched noise segments of the same noise type in training and testing as a function of the number of GMMs in the classification stage. A high complexity classifier of 256 GMMs employed in Kim et al. (2009) was able to learn all spectro-temporal characteristics of the noise, when the same short noise segment was used in training and testing. This was due to an over-fitting of the segregation system which resulted in high H - FA rates (May and Dau, 2014b) and potentially explains the high intelligibility scores obtained in the study. In Experiment I, these observations from May and Dau (2014b) were verified. The measured intelligibility scores of the front-end strategy were higher with 64 GMMs in Fig. 2.3b compared to the lower complexity classifier of 16 GMMs in Fig. 2.3a. Employing the same amount of components as in Kim et al. (2009) would likely result in intelligibility scores at ceiling and close to the IBM.

The ability of segregation systems to generalize to acoustic conditions not seen during training is a very important aspect. In Experiment II novel noise segments in testing not seen during training were considered. Despite

that intelligibility increased with appended subbands in Fig. 2.6a, none of the configurations were able to improve speech intelligibility over noisy speech, suggesting that the system ability to generalize to unseen segments of the six-talker noise was limited. This noise type contains spectro-temporal modulations very similar to modulations in the target speech signal. Therefore, the task of improving intelligibility in a realistic setup is non-trivial. According to May and Dau (2014b), the H - FA rates were generally lower when the considered segregation system was tested with unseen noise segments of the same noise recording, and the rates decrease with increasing GMM classifier complexity. Therefore, in a more realistic setup like in Experiment II, choosing a lower complexity classifier will reduce the risk of over-fitting the system (May and Dau, 2014b), however at the expense of lower H - FA rates and lower intelligibility outcomes.

Other studies have successfully demonstrated a system ability to generalize well to acoustical mismatches by employing DNNs because of their predictive power and the ability to benefit from large-scale training for feature learning (Healy et al., 2015; Chen et al., 2016a,b). In Healy et al. (2015), a 4-hidden layer DNN was applied and tested on novel segments of the same noise type which led to a 25% improvement in WRS in 20-talker babble at −5 dB SNR in NH listeners, but no improvement in cafeteria noise. In Chen et al. (2016a), a multi-conditional training set was introduced, and a classifier was trained using a 5-hidden layer DNN and tested for a range of novel noise types. For the same 20-talker noise at −5 dB SNR, they were able to improve the WRS by approximately 10% in NH listeners. The amount of training employed in these two studies, however, differs from the current study. In Healy et al. (2015) $560 \times 50 = 28,000$ utterances were used for each noise type and SNR, and in Chen et al., 2016a 640,000 utterances were used in the multi-conditional training set. In the current study, only 210 utterances were used for training of the GMM classification stage. The capability of the DNNs to handle large-scale training data is most likely key to an increased ability to generalize to the unseen acoustical conditions.

2.5.3 Implications for cost function design

Kressner et al. (2016) highlighted potential limitations of STOI in predicting the intelligibility of binary-masked speech. In the present study, ESTOI was

employed instead of STOI, but several observations indicated that ESTOI has similar limitations as STOI. First of all, in Experiment I the ranking of the system configurations for the ESTOI improvements conflicted with the ranking of the configurations for the measured intelligibility improvements, as was observed in Fig. 2.4. Secondly, in Experiments I and II, ESTOI predicted improvements of the system configurations when no intelligibility improvements were actually present. In Experiment I, the listener study only revealed improvements for configurations with the 64 GMMs classifier, and in Experiment II no improvements were observed at all. Therefore, ESTOI alone is not able to account for the observations in this study. Furthermore, the H - FA metric was also not able to correctly predict the ranking of the system configurations in Experiment I. Specifically, the H - FA rate was consistently higher for the back-end strategy than the front-end strategy, despite that the intelligibility study revealed an opposite effect. Therefore, it is possible to construct a segregation system that is able to improve H - FA and ESTOI, but, at the same time, fails to improve speech intelligibility scores in noisy conditions. This reveals the limitations of the two measures and emphasizes the need of a single objective measure that comprehensively predicts segregation performance and correlates well with intelligibility for speech segregation systems.

The findings from Experiment I and II have important implications for the design of cost functions in computational speech segregation systems. Monitoring the amount of mask clustering γ in the estimated IBMs seems critical as the clustering among erroneously-labeled T-F units should be minimized. The IBM itself inherently contains clustering, and the obtained γ value can be regarded as the accepted amount of clustering among the correctly-labeled T-F units. Therefore, an appropriate cost function should maximize the H - FA rate and approximate γ as close as possible to γ of the IBM.

2.6 Conclusion

In this study, two experiments were conducted with NH listeners. In Experiment I, the impact of spectro-temporal context in a computational speech segregation system was investigated by considering two strategies in the system front-end and back-end, respectively. The experiment showed that computing delta features in the front-end led to higher speech intelligibility than employing

an SVM-based integration strategy in the back-end. The results were consistent across different noise types and for different classifier complexities. In Experiment II, the delta features were explored in detail and tested in a setup that considered novel noise segments of the same six-talker noise. Intelligibility scores increased with the amount of spectral information exploited, but the segregation system failed to generalize to novel noise segments of this particular noise type. The intelligibility scores were subsequently compared to predictions from several objective measures. The comparison showed that no single measure could account for all intelligibility scores, and therefore emphasizes the need of a single objective measure that comprehensively predicts segregation performance and correlates well with intelligibility. The findings from the present study may have implications for the design of computational speech segregation systems, in which spectro-temporal context should be incorporated without increasing the amount of clustering among erroneous labeled T-F units. Furthermore, the findings can help select a cost function that correlates with intelligibility. According to the results in the present study, the cost function should maximize the H - FA rate and approximate the γ value as close as possible to the γ of the IBM.

Acknowledgments

This work was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences, the EU FET grant TWO!EARS, ICT-618075 and by the Danish Council for Independent Research (DFF) with grant number DFF-5054-00072. Part of this work was presented at the 17th Annual Conference of the International Speech Communication Association, San Francisco, USA (Bentsen et al., 2016).

3

The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility [a]

Abstract

Computational speech segregation attempts to automatically separate speech from noise. This is challenging in conditions with interfering talkers and low signal-to-noise ratios. Recent approaches have adopted deep neural networks (DNNs) and successfully demonstrated speech intelligibility improvements. A selection of components may be responsible for the success with these state-of-the-art approaches: the system architecture, a time frame concatenation technique and the learning objective. The aim of this study was to explore the roles and the relative contributions of these components by measuring speech intelligibility in normal-hearing listeners. A substantial improvement of 25.4 percentage points in speech intelligibility scores was found going from a subband-based architecture, in which a Gaussian Mixture Model-based classifier predicts the distributions of speech and noise for each frequency channel, to a state-of-the-art DNN-based architecture. Another improvement of 13.9 percentage points was obtained by changing the learning objective from the ideal binary mask, in which individual time-frequency units are labeled as either speech- or noise-dominated, to the ideal ratio mask, where the units are assigned a continuous value between zero and one. Therefore, both components play

[a] This chapter is based on: Bentsen, T., T. May, A. A. Kressner, and T. Dau (2018). The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility. *PloS ONE* 13 (5): e0196924. https://doi.org/10.1371/journal.pone.0196924.

significant roles and by combining them, speech intelligibility improvements were obtained in a six-talker condition at a low signal-to-noise ratio.

3.1 Introduction

Computational speech segregation attempts to automatically separate speech from interfering noise. This is particularly challenging in single-channel recordings where a speech signal is corrupted by competing talkers and the signal-to-noise ratio (SNR) is low. It has been suggested to exploit *a priori* knowledge about the speech signal and the interfering noise by constructing an ideal binary mask (IBM) (Wang, 2005). Specifically, the IBM is derived by comparing the SNRs in individual time-frequency (T-F) units to a local criterion (LC). The resulting IBM consists of binary values where T-F units with SNRs exceeding the LC are considered to be speech-dominated and labeled one, whereas T-F units with SNR below the LC are considered to be noise-dominated and are labeled zero. However, since the IBM is unavailable in realistic scenarios, the challenge in computational speech segregation is to estimate the IBM from the noisy speech. Typically, computational speech segregation systems consist of an acoustic feature extraction stage combined with a classification stage where the T-F units are separated into speech-dominated and noise-dominated units in the estimated mask.

In many studies, objective measures have been used to optimize the performance of computational speech segregation systems during the development stage. One commonly used objective measure has been the H - FA rate, which calculates the difference between the percentage of correctly classified speech-dominated T-F units (hit rate, H) and the percentage of incorrectly classified noise-dominated T-F units (false alarm rate, FA) (Kim et al., 2009; Han and Wang, 2012; May and Dau, 2013; Wang and Wang, 2013; May and Dau, 2014a; May and Dau, 2014b; May et al., 2015). Another commonly used objective measure has been the short-term objective intelligibility (STOI) (Taal et al., 2011; Wang et al., 2014; Jensen and Taal, 2016; Zhang and Wang, 2016). However, both objective measures have limitations in predicting speech intelligibility. This has been observed with configurations in which the IBM has been degraded with different mask errors (Kressner et al., 2016), and with computational speech segregation systems for noise reduction (Gelderblom

et al., 2017; Bentsen et al., 2018b). Measuring speech intelligibility in listeners is therefore important to properly evaluate changes introduced in a speech segregation system.

Recent approaches in computational speech segregation have considered systems in which the T-F units are predicted by deep neural networks (DNNs). With these *state-of-the-art approaches*, measured speech intelligibility improvements have been demonstrated in various adverse conditions (Healy et al., 2015; Chen et al., 2016a; Healy et al., 2017; Kolbæk et al., 2017). A selection of components may be responsible for the success: the system architecture, a time frame concatenation technique and the learning objective.

First, the system architecture is different than in previously used approaches. In the state-of-the-art approaches, the features are extracted per frequency channel and subsequently stacked across frequency. The T-F units in the estimated mask are then predicted simultaneously across all frequency channels by the DNN. This has consequences for how the context, i.e. the spectro-temporal regions in the estimated mask where speech-dominated T-F units tend to cluster, is exploited by the system. By predicting the T-F units simultaneously across all frequency channels, the state-of-the-art approaches therefore exploit the spectral context in a broadband manner. In previously used approaches, a classifier has been employed per frequency channel (i.e., a subband classifier) in the speech segregation system. These subband classifiers have been implemented with either Gaussian mixture models (GMMs) (Kim et al., 2009), support vector machines (SVMs) (Han and Wang, 2012; Wang and Wang, 2013) or DNNs (Healy et al., 2013). In such a subband-based system, the spectral context has been exploited across neighboring subbands by, for example, including delta features which can capture spectral feature variations across neighboring frequency channels (Kim et al., 2009; May et al., 2015; Bentsen et al., 2016).

Secondly, state-of-the-art approaches often exploit temporal context by concatenating extracted feature vectors across a predefined number of time frames (Wang et al., 2014; Chen et al., 2016a; Zhang and Wang, 2016). Past and future time frames have both been considered. Improvements in objective measures with time frame concatenation have been reported (Wang et al.,

2014). However, the effect of employing a time frame concatenation technique on measured speech intelligibility is currently unknown.

Thirdly, state-of-the-art approaches use the ideal ratio mask (IRM) as the learning objective instead of the IBM (Healy et al., 2015; Chen et al., 2016a,b; Healy et al., 2017; Kolbæk et al., 2017). In the IRM, the mask value is a continuous gain between zero and one and computed according to the *a priori* SNR of the considered T-F unit (Srinivasan et al., 2006; Narayanan and Wang, 2013; Hummersone et al., 2014; Wang et al., 2014). Therefore, the IRM is similar to an ideal Wiener filter (Hummersone et al., 2014). The perceptual effect of applying IBMs versus IRMs has been investigated in terms of speech quality (Brons et al., 2012). A higher sound quality rating with lower noise annoyance and a larger degree of speech naturalness were observed when using IRMs compared to IBMs. Additionally, continuous versus binary gain functions were compared in the framework of minimum mean-squared error (MMSE)-based noise reduction algorithms (Jensen and Hendriks, 2012). It was shown that the continuous gain function outperformed the binary gain function in terms of measured speech intelligibility scores. Furthermore, a larger STOI improvement relative to noisy speech was found with IRM estimation in DNN-based systems compared to IBM estimation (Wang et al., 2014; Zhang and Wang, 2016). Despite these observations, none of the state-of-the-art approaches has actually demonstrated measured speech intelligibility improvements with IRM estimation over IBM estimation in an otherwise identical system. Furthermore, it is unclear how much IRM estimation contributes to the success of state-of-the-art approaches, especially in comparison to the other components.

The aim of the present study was to explore the roles and the relative contributions of these components within state-of-the-art computational speech segregation by measuring speech intelligibility in normal-hearing (NH) listeners at a low SNR. Specifically, a broadband DNN-based system was compared with a corresponding subband-based system. The subband-based system employed a GMM classifier per frequency channel using delta features across subbands to exploit the spectral context. To exploit temporal context in the DNN-based system, time frame concatenation was either included or excluded. Moreover, the effect of IRM estimation versus IBM estimation was studied in the DNN-

based system. To create as fair of a comparison between the different systems as possible, the DNN-based system and the subband GMM-based system considered the same features, and were both trained using the same amount of training data. Therefore, the considered systems were not necessarily designed to maximize the measured speech intelligibility, but instead are designed to be able to systematically compare each of the different components.

3.2 Methods

3.2.1 Feature extraction

Noisy speech was sampled at a rate of 16 kHz and decomposed into $K = 31$ frequency channels by employing an all-pole version of the gammatone filterbank (Lyon, 1997), whose center frequencies were equally spaced on the equivalent rectangular bandwidth (ERB) scale between 80 and 7642 Hz. Previous studies (Kim et al., 2009; May and Dau, 2014a; May et al., 2015) successfully exploited modulations in the speech and the interferer by extracting amplitude modulation spectrogram (AMS) features (Kollmeier and Koch, 1994; Tchorz and Kollmeier, 2003). To derive the AMS features in each frequency channel (subband), the envelope was extracted by half-wave rectification and low-pass filtering with a cutoff frequency of 1 kHz. Then, each envelope was normalized by its median computed over the entire envelope signal. These normalized envelopes were then processed by a modulation filterbank that consisted of one first-order low-pass and five band-pass filters with logarithmically spaced center frequencies and a constant Q-factor of 1. The cutoff frequency of the modulation low-pass filter was set to the inverse of the window duration to ensure that at least one full period of the modulation frequency was included in the window (May et al., 2015). Using time frames of 32 ms with 75 % overlap (i.e., a 8 ms frame shift) resulted in a cutoff frequency of 32 Hz. The root mean square (RMS) value of each modulation filter was then calculated across each time frame.

3.2.2 The DNN-based system

Figure 3.1 illustrates the DNN-based system. The AMS feature space was power-compressed with an exponent of 1/15 (Chen et al., 2016a), stacked across frequency channels and fed to the input layer of a feed-forward DNN. The

network architecture consisted of an input layer, two hidden layers that each had 128 nodes, and an output layer of 31 nodes. Feature frame concatenation was employed by appending the five past AMS feature time frames to the current frame, which corresponded to a temporal context of 40 ms. The DNN-based system was used to either estimate the IBM or the IRM. The IRM was given by (Wang et al., 2014):

$$\mathrm{IRM}(t,f) = \left(\frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)}\right)^{\beta} = \left(\frac{\mathrm{SNR}(t,f)}{\mathrm{SNR}(t,f)+1}\right)^{\beta} \tag{3.1}$$

In Eq. (3.1), the $S^2(t,f)$ and the $N^2(t,f)$ indicate the speech and noise energies, respectively, in a given T-F unit with time frame t and frequency channel f, and β denotes the mask exponent. Mask values in the IRM are therefore scaled according to the SNR, such that T-F units with lower SNR are attenuated more strongly.

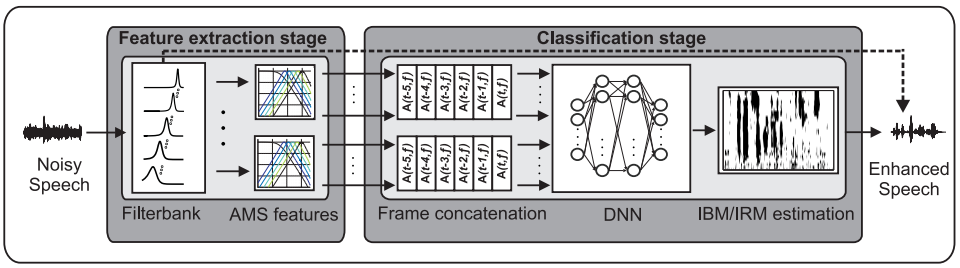


Figure 3.1: Noisy speech was decomposed by a gammatone filterbank and AMS features were extracted per subband. The AMS features were fed into an DNN with two hidden layers of 128 nodes each. The system estimated a time-frequency mask (either an IBM or an IRM), and the mask was subsequently applied to the subband signals of the noisy speech, as illustrated by the dashed line, in order to reconstruct the speech signal.

3.2.3 The subband-based system

The subband-based system has previously been employed (May et al., 2015; Bentsen et al., 2016, 2018b) and a detailed description is given in Bentsen et al. (2018b). In short, delta features were computed symmetrically across frequency bands, resulting in the feature vector $\mathbf{X}(t,f)$:

$$\begin{aligned}
\mathbf{X}(t,f) &= \left[\mathbf{A}(t,f), \Delta_{f-k}\mathbf{A}(t,f), \Delta_{f+k}\mathbf{A}(t,f)\right] \\
\Delta_{f-k}\mathbf{A}(t,f) &= \mathbf{A}(t,f) - \mathbf{A}(t,f-k), \quad \forall k \in \{n \in [1;K] | f-n \geq 1\} \\
\Delta_{f+k}\mathbf{A}(t,f) &= \mathbf{A}(t,f) - \mathbf{A}(t,f+k), \quad \forall k \in \{n \in [1;K] | f+n \leq K\}
\end{aligned} \tag{3.2}$$

In Eq. (3.2), f indicates the current subband and k the considered number of subbands across which the delta features were computed. Seven subbands ($k = 3$) were used in this comparison, since having more than seven subbands does not statistically improve the measured speech intelligibility scores Bentsen et al., 2018b. The classification back-end consisted of a GMM classifier trained to represent the speech and noise-dominated AMS feature distributions ($\lambda_{1,f}$ and $\lambda_{0,f}$) for each subband f of the K filters (Kim et al., 2009). To separate the feature vector into speech- and noise-dominated T-F units, an LC was applied to the *a priori* SNR. The GMM classifier output was given as the posterior probability of speech and noise presence $P\left(\lambda_{1,f}|\mathbf{X}\left(t,f\right)\right)$ and $P\left(\lambda_{0,f}|\mathbf{X}\left(t,f\right)\right)$, respectively:

$$P\left(\lambda_{1,f}|\mathbf{X}\left(t,f\right)\right) = \frac{P\left(\lambda_{1,f}\right)P\left(\mathbf{X}\left(t,f\right)|\lambda_{1,f}\right)}{P\left(\mathbf{X}\left(t,f\right)\right)} \tag{3.3}$$

$$P\left(\lambda_{0,f}|\mathbf{X}\left(t,f\right)\right) = \frac{P\left(\lambda_{0,f}\right)P\left(\mathbf{X}\left(t,f\right)|\lambda_{0,f}\right)}{P\left(\mathbf{X}\left(t,f\right)\right)} \tag{3.4}$$

The *a priori* probabilities $P\left(\lambda_{1,f}\right)$ and $P\left(\lambda_{0,f}\right)$ were computed by counting the number of feature vectors for each of the classes $\lambda_{1,f}$ and $\lambda_{0,f}$ during training.

3.2.4 System configurations

In this study, six system configurations were compared (see Table 3.1). System configurations "GMM (IBM, 1 subband)" and "GMM (IBM, 7 subbands)" exploited spectral context in the subband-based system. In the "GMM (IBM, 1 subband)" configuration, delta features were used as in Kim et al. (2009) with only the adjacent subband. In the "GMM (IBM, 7 subbands)" configuration, $k = 3$ symmetrically placed subbands around the considered subband were used to exploit spectral context, according to Eq. (3.2). Configurations "DNN (IBM)", "DNN (IBM, 40 ms)", "DNN (IRM)" and "DNN (IRM, 40 ms)" were all configurations of the DNN-based system. "DNN (IBM)" and "DNN (IRM)" were configurations with no frame concatenation and using IBM and IRM estimation, respectively. "DNN (IBM, 40 ms)" and "DNN (IRM, 40 ms)" were configurations with five past concatenated frames corresponding to 40 ms duration, and with IBM and IRM estimation, respectively. In addition to the six system configurations, unprocessed noisy speech was tested as a baseline.

Table 3.1: Overview of the system configurations.

Configuration	Classifier	Architecture	Frame concatenation	Learning objective
GMM (IBM, 1 subband)	GMM	Subband	-	IBM
GMM (IBM, 7 subbands)	GMM	Subband	-	IBM
DNN (IBM)	DNN	Broadband	0 ms	IBM
DNN (IBM, 40 ms)	DNN	Broadband	40 ms	IBM
DNN (IRM, 40 ms)	DNN	Broadband	40 ms	IRM
DNN (IRM)	DNN	Broadband	0 ms	IRM

3.2.5 Stimuli

The speech material was taken from the Danish Conversational Language Understanding Evaluation (CLUE) database (Nielsen and Dau, 2009). It consists of 70 sentences in 7 lists for training and 180 sentences in 18 balanced lists for testing, and the sentences are spoken by a male Danish talker. Noisy speech mixtures were created by mixing individual sentences with the non-stationary six-talker (ICRA7) noise (Dreschler et al., 2001). A Long Term Average Spectrum (LTAS) template was computed based on the CLUE corpus, and the LTAS of the noise masker was adjusted to the template LTAS. A randomly-selected noise segment was used for each sentence. In order to avoid onset effects in the speech intelligibility test (Nielsen and Dau, 2009), the noise segment started 1000 ms before the speech onset and ended 600 ms after the speech offset.

3.2.6 System training and evaluation

The full ICRA7 noise recording of 600 s was divided such that one half of the recording was used for training and the other half was used for testing. The 70 training sentences were each mixed three times with a randomly-selected noise segment from the noise recording at $-5, 0$, and 5 dB SNR to create a training set of 210 utterances. Training at multiple SNR has been used as an approach in many studies, e.g. in Kim et al. (2009). This training set was used to train both the DNN-based system and the subband GMM-based system. The DNN was trained to estimate either the IBM or the IRM using back-propagation with the scaled conjugate gradient algorithm and a mean-squared error cost function. All hidden layers were trained simultaneously in the network. For the IRM estimation, β was set to 0.5 as previously done (Wang et al., 2014; Zhang and Wang, 2016). For the subband GMM-based system, a moderate classifier complexity of 16 Gaussian components with full covariance matrix was selected. The classifiers were first initialized by 15 iterations of the K-

means clustering algorithm, followed by five iterations of the expectation-maximization algorithm, and an LC of −5 dB was employed. Both systems were evaluated with 180 CLUE sentences that were each mixed with ICRA7 noise at −5 dB SNR.

3.2.7 Subjects and experimental setup

The experiment was conducted with a group of 20 NH listeners that were aged between 20 and 32 years with a mean of 24.5 years. Requirements for participation were: (1) aged between 18 − 40 years, (2) audiometric thresholds of less than or equal to 20 dB hearing level (HL) in both ears (between 0.125 and 8 kHz), (3) Danish as their native language, and (4) no previous experience with the Hearing In Noise Test (HINT) (Nielsen and Dau, 2011) or CLUE material (Nielsen and Dau, 2009).

The total session lasted about two hours, including the screening process. The experiment was approved by the Danish Science-Ethics Committee (reference H-16036391). Listeners were recruited with online advertisement, and they were paid for their participation. Informed consent was obtained prior to the experiment. The subjects were all recruited and tested within a two-month period. The experiment was split into two parts: subject training and subject testing. In the training part, five randomly selected sentences from the training set were presented for each of the conditions to familiarize the subjects with the task. Subsequently, each subject heard one list per condition, whereby conditions and lists were randomized across subjects. The sentences were presented diotically to the listener via headphones (Sennheiser HD650) in an acoustically and electrically shielded booth. Prior to the actual experiments, the headphones were calibrated by first adjusting to a reference sound pressure level (SPL) and then performing a headphone frequency response equalization. During the experiment, the sentences were adjusted to the desired presentation level, and the equalization filters were applied. The SPL was set to a level of 65 dB. For each sentence, the subjects were instructed to repeat the words they heard, and an operator scored the correctly understood words via a MATLAB interface. The subjects were told that guessing was allowed. They could listen to each sentence only once, and breaks were allowed according to the subject's preference.

3.2.8 Statistical analysis

Intelligibility scores were reported as a percentage of correctly scored words, i.e. the word recognition score (WRS). The WRSs were computed per sentence and averaged across sentences per list. The intelligibility scores followed a normal distribution, and a linear mixed effect model was constructed with list WRSs as the response variable and the system configurations as a fixed factor (8 levels).Subjects were treated as a random factor, as is standard in a repeated measures design. Fixed factor levels were tested at a 5% significance level. To visualize the data, the predicted least-squares means and 95% confidence limits of the least-squares means were extracted from the model. To assess any difference between system configurations, the differences of the least-squares means were computed in pairwise comparisons, where the p values were adjusted following the Tukey multiple comparison testing. To evaluate potential speech intelligibility improvements, Paired Student's t-tests between the noisy speech and the relevant system configuration was constructed and tested at a 5% significance level.

3.3 Results

Figure 3.2 shows the measured WRSs of the six system configurations along with unprocessed noisy speech. The sample mean across subjects and a 95% Student's t-based confidence interval of the sample mean were computed and included in Fig. 3.2 for visualization. For the six system configurations, the least-squares means and 95% confidence limits of the least-squares means predictions are shown. In noisy speech, the average WRS was 65%. The relatively high baseline score was presumably due to the fluctuations in the six-talker noise, which has been shown to facilitate listening-in-the-dips in NH subjects Festen and Plomp, 1990. Measured WRSs increased significantly from the "GMM (IBM, 1 subband)" configuration to the "GMM (IBM, 7 subbands)" configuration by 18.9 percentage points ($p < 0.0001$). This result indicates that an increased number of appended delta feature vectors across frequency in the subband GMM-based system led to higher measured speech intelligibility, since a larger amount of spectral context was exploited. Comparing across the systems, the "DNN (IBM)" configuration led to 25.4 percentage points higher WRS than the "GMM (IBM, 1 subband)" configuration ($p < 0.0001$). Despite

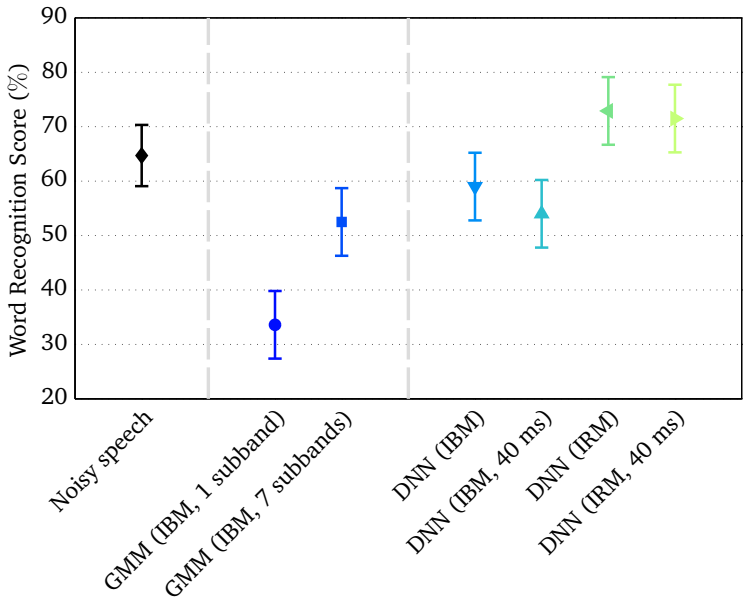


Figure 3.2: Unprocessed noisy speech served as a baseline condition. For the baseline (diamonds), sample means across subjects and 95% Student's t-based confidence intervals of the mean were computed. For the system configurations, the least-squares means and 95% confidence limits of the least-squares means predictions derived from the linear mixed effect model were plotted.

that the "DNN (IBM)" configuration had a higher WRS of 6.5 percentage points than the "GMM (IBM, 7 subbands)" configuration, measured speech intelligibility scores for the two configurations were not significantly different.

The "DNN (IBM)" and "DNN (IBM, 40 ms)" configurations did not differ significantly from each other, and no statistically significant difference was found either between the "DNN (IRM)" and "DNN (IRM, 40 ms)" configurations. Therefore, the employed time frame concatenation technique, which was used to exploit temporal context, did not have a perceptual effect in the current DNN-based system, regardless of whether IBM or IRM estimation was considered in the system.

The configuration "DNN (IRM)" led to 13.9 percentage points higher WRS than the "DNN (IBM)" configuration ($p < 0.001$). Furthermore, 17.5 percentage points higher WRS was observed for the "DNN (IRM, 40 ms)" configuration

than for the "DNN (IBM, 40 ms)" configuration ($p < 0.0001$). Therefore, a clear perceptual advantage was found for IRM over IBM estimation in the DNN-based system. The measured intelligibility scores were subsequently converted into WRS improvements relative to the unprocessed noisy speech. Significant improvements, based on the Paired Student's t-tests at a 5% significance level, were obtained for the "DNN (IRM)" configuration (8.2 percentage points; $t[19] = 2.36; p = 0.014$) and the "DNN (IRM, 40 ms)" configuration (6.8 percentage points; $t[19] = 2.14; p = 0.023$). This particular finding demonstrates the benefit of estimating the IRM as opposed to the IBM, when computational speech segregation systems are used for noise reduction applications.

3.4 Discussion

3.4.1 The roles and relative contributions of the components

The comparison between the subband GMM-based system configurations ("GMM (IBM, 1 subband)" and "GMM (IBM, 7 subbands)") indicated that the measured speech intelligibility scores increased with the number of subbands used to compute the delta features. By increasing the number of subbands, the AMS feature vector was rapidly growing. In Bentsen et al. (2018b), it was shown that more than seven considered subbands did not further increase the measured speech intelligibility. The subband GMM classifier was therefore limited in the capability to handle the large amount of AMS feature data. In addition, the "GMM (IBM, 1 subband)" configuration that resembled previously-used approaches (Kim et al., 2009; May et al., 2015; Bentsen et al., 2016) resulted in a much lower speech intelligibility than the corresponding broadband DNN-based system configuration ("DNN (IBM)"). By increasing the number of subbands and thereby exploiting more spectral context in the subband GMM-based system, it was possible to achieve a measured speech intelligibility score similar to that obtained with the DNN-based system. By changing the architecture from subband GMM classifiers to a broadband DNN, the segregation system was able to predict the T-F units simultaneously across all of the subbands. Therefore, the DNN-based system exploited the spectral context in a broadband manner, which may be more effective than the corresponding subband-based system. This is most likely because of the capability of DNNs to handle higher-dimensional feature vectors. Estimated

IBMs with these three configurations ("GMM (IBM, 1 subband)", "GMM (IBM, 7 subbands)" and "DNN (IBM)") are shown in Figs. 3.3f-3.3h and can be compared to the IBM in Fig. 3.3e. H-FA rates were computed for each of the estimated IBMs to indicate the mask estimation accuracy. Results were 27.8% ("GMM (IBM, 1 subband)"), 34.5% ("GMM (IBM, 7 subbands)") and 63.7% ("DNN (IBM)"), respectively. A larger amount of spectral context is exploited by increasing the number of considered subbands in the subband GMM-based system (Figs. 3.3f and 3.3g), which leads to more correctly-classified speech T-F units (hits) and therefore a larger H-FA rate. However, the estimated IBM using the DNN-based system (Fig. 3.3h) contains much larger regions with correctly-classified speech T-F units and and less mask errors (both misses and false alarms), which has increased the H-FA rate quite substantially. The results of the present study also indicated that the employed time frame concatenation technique, which has been proposed to exploit temporal context in the state-of-the-art approaches (Wang et al., 2014; Chen et al., 2016a; Zhang and Wang, 2016), did not have a significant impact on the measured speech intelligibility. This was observed regardless of whether the DNN-based system estimated the IBM or the IRM. This result was rather surprising, but should be seen in light of the small amount of training data (only 210 utterances) fed to the DNN-based system. Most likely, the small amount of training data was not sufficient to unfold the predictive power of the DNN. Another important point is that "only" five past feature frames were appended to the current frame, resulting in an exploited temporal context of 40 ms. To put this into perspective, 23 frames were concatenated in total with a step size of 10 ms in another study (Chen et al., 2016a), which resulted in a much larger exploited temporal context of 200 ms. Furthermore, the 23 frames were symmetrically placed around the current frame with eleven past and eleven future time frames. Whether the temporal context in future time frames affect speech intelligibility is not clear.

A substantial perceptual advantage of IRM over IBM estimation was observed in the DNN-based system, where both configurations with IRM estimation ("DNN (IRM)" and "DNN (IRM, 40 ms)") led to higher measured speech intelligibility scores than the corresponding configurations with IBM estimation. The present study therefore demonstrated the effectiveness of the IRM estimation over the IBM estimation with respect to measured speech

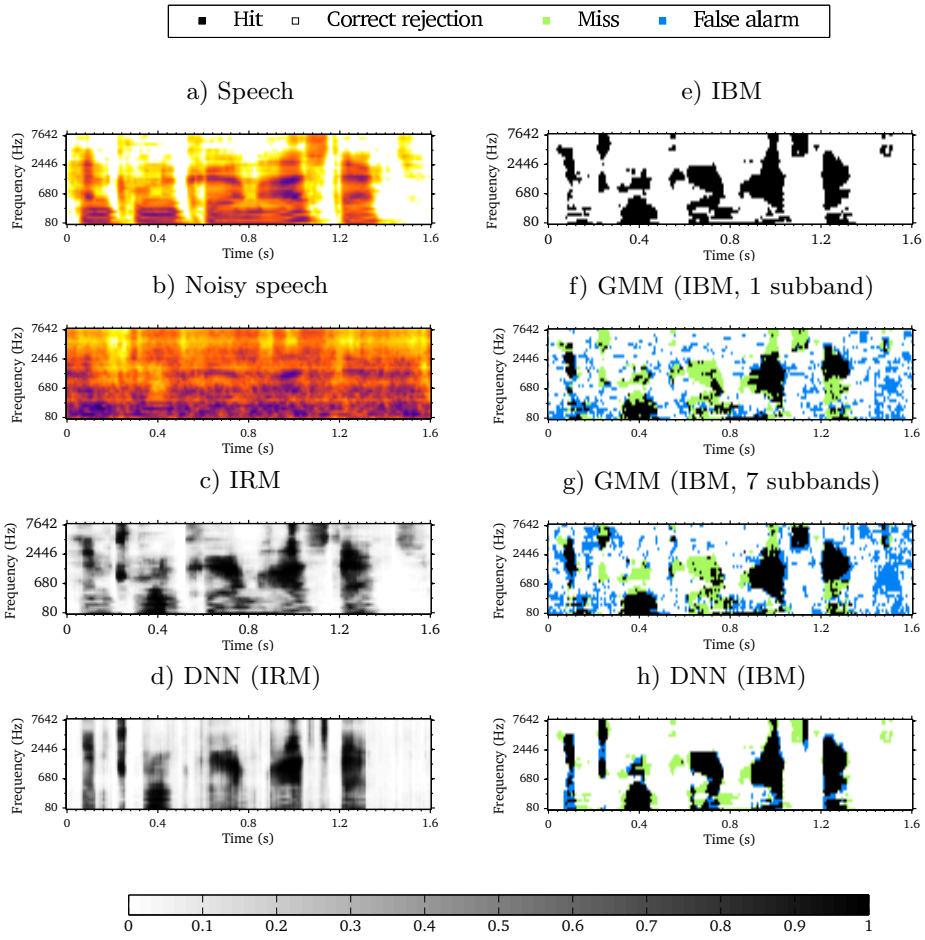


Figure 3.3: Estimated and ideal time-frequency masks for an CLUE sentence mixed with ICRA7 noise at −5 dB SNR. The spectrograms of clean and noisy speech are shown in Figs. 3.3a and 3.3b. The IRM and the IBM are shown in Figs. 3.3c and 3.3e. A selection of estimated masks from system configurations are shown in Figs. 3.3d, 3.3f, 3.3g and 3.3h. Misses (speech-dominated T-F units erroneously labeled as noise-dominated) and false alarms (noise-dominated T-F units erroneously labeled as speech-dominated) are shown on top of the estimated IBMs. The estimated IBM in Fig. 3.3h was converted from the corresponding estimated IRM by applying a threshold, which was derived from Eq. (3.1) at −5 dB SNR and using $\beta = 0.5$.

intelligibility in the state-of-the-art approaches. The effectiveness of the IRM can be explained by how the mask gain values are computed. From Eq. (3.1), it is observed that these values can vary continuously between 0 and 1. Comparing the ideal masks (Figs. 3.3c and 3.3e) to the spectrogram of speech in quiet (Fig. 3.3a), it can be seen that several mask regions with low speech energy are captured by the IRM, but not by the IBM (e.g., around

0.6 s and above 2446 Hz). The IRM can therefore convey important speech information that is not reflected in the IBM, suggesting that the IRM is a better learning objective than the IBM in computational speech segregation. By comparing the estimated masks in Figs. 3.3d and 3.3h, it is also apparent that the estimated IRM mask values are more tolerant to misses by the segregation system. Several mask regions with misses in Fig. 3.3h correspond to areas with positive gain values in Fig. 3.3d, such that speech information is conveyed, which otherwise would have been missed. Therefore, even though a binary classification of T-F units makes the IBM a simpler objective to estimate, the findings in the present study support the use of the IRM estimation in state-of-the-art approaches for noise reduction applications. In addition to the measured speech intelligibility, subjective speech quality will most likely also improve with IRM estimation, since it has previously been demonstrated that the IRM itself improves the quality in comparison to the IBM (Brons et al., 2012).

Finally, the relative contributions of the components within state-of-the-art approaches were addressed. First, a substantial improvement of 25.4 percentage points in measured speech intelligibility scores was found by changing the system architecture from subband GMM-based, with first-order delta features across frequency, to the broadband DNN architecture. The subband GMM-based architecture was similar to previously-used system architectures (Kim et al., 2009; May et al., 2015; Bentsen et al., 2016). Secondly, by changing from IBM estimation to IRM estimation, another improvement of 13.9 percentage points in measured speech intelligibility scores was obtained. Therefore, these results suggest that both of these components play a significant role in the success of the state-of-the-art approaches. By combining the two significant components, intelligibility improvements of about $7-8$ percentage points relative to noisy speech were demonstrated. These improvements were obtained despite that the system was evaluated in the challenging scenario of being presented with unseen, six-talker noise at a low SNR after a relatively limited system training.

Large-scale training in the DNN-based system

Being able to generalize to acoustic conditions not seen during training (i.e., mismatches between acoustic conditions encountered during training and testing) is crucial for any speech segregation system to be applied in realistic scenarios. The segregation systems in this study considered a mismatch of

six-talker noise segments between training and testing. One reason for the relatively limited speech intelligiblity improvement with the DNN-based system with IRM estimation, in comparison to that which has been reported in other studies, is that the competing six-talker noise contains spectro-temporal modulations that are very similar to the modulations in the speech signal. This complicates the task of automatically segregating the interfering noise from the target speech. Other studies have demonstrated a generalization ability with DNN-based systems but have employed 20-talker noise with less fluctuations (Healy et al., 2015; Chen et al., 2016a).

Another reason for the limited improvement is the small amount of training data used in the present study. The training set was kept low with only 210 utterances in order to compare the DNN-based system with the subband GMM-based system. However, it has previously been shown that DNNs can benefit from large-scale training in computational speech segregation (Chen et al., 2015, 2016a,b), and intelligibility improvements over noisy speech can be obtained with these systems in conditions with various acoustic mismatches (Healy et al., 2015; Chen et al., 2016a; Healy et al., 2017; Kolbæk et al., 2017). In one of these studies (Healy et al., 2015), the speech segregation system was trained with 28,000 utterances presented in different types of noise at different SNRs. At −5 dB SNR and with 20-talker noise, this led to an improvement of 25 percentage points in speech intelligibility scores in NH listeners. In another study (Chen et al., 2016a), the system was trained with 640,000 utterances in a multi-conditional training set to produce an improvement of 10 percentage points in the speech intelligibility scores in the same experimental design as the first study (Healy et al., 2015). Retraining the considered DNN-based system with a larger training set than 210 utterances would most likely improve the generalization ability to the unseen six-talker noise segments. Large-scale training is therefore also an important component within state-of-the-art approaches in computational speech segregation, and investigating the impact of large-scale training on measured speech intelligibility is one direction for future work.

3.5 Conclusion

This study explored the relative contributions of a selection of components within state-of-the-art speech segregation systems to improving speech intelli-

gibility. The first component was the system architecture, which was changed from subband-based, in which a classifier was employed per frequency channel, to a DNN network architecture where the T-F units were predicted simultaneously across all frequency channels. Specifically, a broadband DNN-based system was compared with a corresponding subband GMM-based system. A second component was the time frame concatenation technique. This technique is often applied in DNN-based speech segregation systems to exploit the temporal context. However, this technique did not show a significant effect on the measured speech intelligibility scores in this study, presumable because of the relatively limited amount of training data was not sufficient to unfold the predictive power of the DNN. The third considered component was the estimation of the IRM instead of estimating the IBM. Results showed a substantial perceptual advantage with the IRM estimation in the DNN-based system. Finally, the relative contributions of the components were addressed. A substantial improvement of 25.4 percentage points in measured speech intelligibility scores was found by changing the system architecture from subband GMM-based, which is similar to previously-used architectures, to a recent DNN architecture. Another improvement of 13.9 percentage points was obtained by changing from IBM estimation to IRM estimation in the state-of-the-art approach. Therefore, both of these components seem to play a significant role in the success of state-of-the-art speech segregation systems. By combining the two significant components, intelligibility improvements of about $7-8$ percentage points relative to noisy speech were demonstrated in adverse conditions where speech was corrupted by a six-talker noise at a low SNR.

Acknowledgments

This work was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences, the EU FET grant TWO!EARS, ICT-618075 and by the Danish Council for Independent Research (DFF) with grant number DFF-5054-00072.

4

The impact of noise power spectral density estimation on speech intelligibility in cochlear-implant speech coding strategies [a]

Abstract

The advanced combination encoder (ACE) is a well-established speech-coding strategy in cochlear-implant (CI) processing that selects a number of frequency channels based on amplitudes. However, speech intelligibility outcomes with this strategy are limited in noisy conditions with low signal-to-noise ratios (SNRs). To improve the speech intelligibility outcome, either noise-dominant channels can be attenuated prior to ACE with noise-reduction strategies or, alternatively, channels can be selected based on estimated SNRs. A noise power spectral density (PSD) estimation stage is, however, required. This study investigated the impact of utilizing an improved noise PSD estimation stage in both noise-reduction strategies and in channel-selection strategies. Results imply that estimation with improved noise-tracking capabilities does not necessarily translate to an increased speech intelligibility when the noise PSD estimation is utilized for noise reduction nor for when it is utilized for channel selection. In addition, the impact of altering the SNR-based channel-selection criterion from fixed to adaptive was investigated. The local criterion (LC) in the adaptive channel-selection is important for reducing the noise-induced stimulation in the CI recipients.

[a] This chapter is based on research in collaboration with Cochlear Limited (Dr Stefan Mauger) during an external research stay at Cochlear Melbourne, Australia. Parts of the chapter have been submitted as a letter to the editor: Bentsen, T., S. Mauger, A. A. Kressner, T. May, and T. Dau (in review). The impact of noise power estimation on speech intelligibility in cochlear-implant speech coding strategies.*J. Acoust. Soc. Am.*, in review.

4.1 Introduction

In cochlear-implant (CI) processing, a signal is decomposed into frequency channels and the advanced combination encoder (ACE) selects a fixed number of channels with the largest amplitudes for electrical stimulation (Wilson et al., 1988; McDermott et al., 1992). However, speech intelligibility outcomes with ACE in noisy conditions with low signal-to-noise ratios (SNRs) are limited primarily because: (i) the channels with the largest amplitudes can be noise-dominated instead of speech-dominated and (ii) ACE always selects a fixed number of channels when the signal amplitude is above a predefined threshold, irrespective of whether speech is present or absent (Hu and Loizou, 2008). In an attempt to improve the speech intelligibility in these noisy conditions, a range of different speech-coding strategies have been developed.

A group of strategies apply noise-reduction prior to stimulation with ACE. Specifically, a noise power spectral density (PSD) estimate is obtained and noise-dominant channels are attenuated before the channels with the largest amplitudes are selected for stimulation. In current CI processors (Dawson et al., 2011; Mauger et al., 2012a,b), noise PSD estimation is based on minimum statistics (MS), where the estimate is obtained by tracking the minimum of the noisy speech PSD in a time window that typically spans over $1-3\,\mathrm{s}$ (Martin, 2001). Substantial speech intelligibility improvements have been demonstrated in speech-weighted noise with noise reduction based on MS-based noise PSD estimators, but the strategy failed to improve speech intelligibility in the presence of more dynamic noises (Mauger et al., 2012a). This may be, at least partly, because the MS-based estimator tracks changes in fluctuating noises with a delay corresponding to the duration of the time window. Since the noise PSD estimate is determined by the minimum within the time window, this can lead to an underestimation of the true noise PSD. Shortening the time window to avoid a large delay and underestimation would increase the likelihood of tracking speech segments instead, since it will be more likely to encounter time windows that do not contain speech gaps. To overcome the limitations of the MS-based estimator, Gerkmann and Hendriks (2012) proposed a PSD estimator based on the speech presence probability (SPP). This noise PSD estimator has been shown to track changes in the true noise PSD faster than the MS-based estimator and has, therefore, been reported to be more accurate than

the MS estimator in terms of the logarithmic estimation error (Gerkmann and Hendriks, 2012). However, whether this improved accuracy translates to higher speech intelligibility in the context of a noise-reduction strategy is not known.

Another group of strategies, which have been proposed to address the low speech intelligibility outcomes in CI recipients in the presence of noise, select which channels to stimulate directly based on an SNR criterion (Hu and Loizou, 2008). A frequency channel with a high instantaneous SNR conveys more reliable speech information than a frequency channel with a low instantaneous SNR, and only channels with high SNRs are therefore selected for stimulation. One approach is to select the n-of-m channels with the highest SNRs. This fixed channel-selection strategy is similar to ACE, except that the channel-selection criterion has changed from amplitude to SNR. Alternatively, a channel is selected only if the SNR is above an local criterion (LC) (Hu and Loizou, 2008). The number of selected channels therefore change adaptively with the SNR, such that in each processing cycle between 0 and m channels are stimulated. With this latter approach, together with *a priori* information of the clean speech and the noise signals to derive the SNR, speech intelligibility has been restored to levels obtained for speech in quiet for both speech-weighted noise and multi-talker babble (Hu and Loizou, 2008; Dawson et al., 2011; Hazrati and Loizou, 2013). The approach has strong similarities with the ideal binary mask (IBM) which is typically used as the learning objective in computational speech segregation (Wang, 2005). In the IBM, the a priori SNR in a specific time-frequency (T-F) unit is compared to an LC to separate the T-F representation of noisy speech into speech-dominated and masker-dominated T-F units. To apply the channel-selection strategies in practice, an SNR estimation algorithm is required. One approach is to consider speech segregation systems that employ machine learning techniques to estimate the SNRs (Hu and Loizou, 2010; Goehring et al., 2017). In Hu and Loizou (2010), a speech segregation system with a high-complexity classifier was trained and tested using the same short noise recording. A high-complexity classifier is able to learn all spectro-temporal characteristics of the noise, if the same short noise recording is used during training and testing. As a result, the system does not generalize well to any mismatches between training and testing (May and Dau, 2014b) and is therefore not feasible in realistic applications. Instead of employing machine-learning based speech segregation systems, the SNRs

can be estimated by using noise PSD estimators. These noise PSD estimators do not require pre-training for a specific acoustical condition, and most of them are real-time applicable because of low latency values. Given the higher accuracy of the SPP-based noise PSD estimator as compared to the MS-based estimator, the algorithm appears to be a promising candidate for this task.

The present study investigated the impact of utilizing the SPP-based noise PSD estimator in a range of noise-reduction and channel-selection strategies by measuring speech intelligibility in CI recipients. In addition, the sound quality was rated by the CI recipients. First, the SPP-based noise PSD estimator was implemented in a noise-reduction strategy, and intelligibility scores were compared to intelligibility scores obtained with the MS-based estimator. Secondly, the estimated SNRs were used in both the fixed and adaptive versions of the channel-selection strategies, and intelligibility scores were compared with intelligibility scores obtained with ACE, as well as with the existing noise-reduction strategy in combination with ACE. With this second set of comparisons, the impact of altering the channel-selection criterion was investigated. At the same time, the effect of the LC in SNR-based channel-selection was evaluated on the speech intelligibility outcome. Specifically, the relative impact of altering the SNR-based channel selection from fixed to adaptive was analyzed.

4.2 Methods

4.2.1 Signal processing

Noisy speech was sampled at 16 kHz and buffered into $\ell = 1...L$ frames of 8 ms duration with 1 ms step size in the CI signal path, shown in Fig. 4.1. A short-time discrete Fourier transform (STFT) with $k = 1...K$ bins ($K = 128$) decomposed the noisy speech in the signal path, and the noise PSD estimate was obtained for each individual STFT bin k. In the following, it is assumed that the speech $S_k(\ell)$ and the noise $N_k(\ell)$ components are complex Gaussian distributed and additive in the STFT domain, such that the noisy component, $Y_k(\ell)$, can be represented as:

$$Y_k(\ell) = S_k(\ell) + N_k(\ell) \tag{4.1}$$

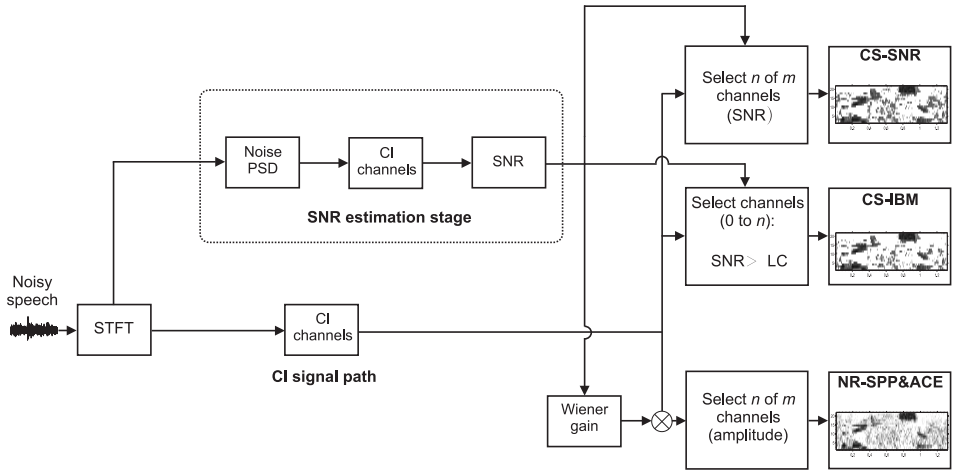


Figure 4.1: Overview of the implemented speech coding strategies. In the CI signal path, noisy speech was first decomposed by an STFT and bins were combined into CI channels. In the SNR estimation stage, a noise PSD was obtained from bins (Gerkmann and Hendriks, 2012) and the estimated noise PSD in each CI channel was used to estimate the a priori SNRs. The a priori SNRs were then used to either select channels ("CS-SNR" and "CS-IBM") or compute a set of gain values which were applied prior to stimulation with ACE ("NR-SPP&ACE').

For a given time frame, ℓ, the current noise PSD estimate, $\widehat{|N_k(\ell)|^2}$, is computed (Gerkmann and Hendriks, 2012):

$$\widehat{|N_k(\ell)|^2} = \Big(1 - P(\mathcal{H}_1|Y_k(\ell))\Big) \cdot |Y_k(\ell)|^2 + P(\mathcal{H}_1|Y_k(\ell)) \cdot \widehat{\sigma}^2_{N,k}(\ell-1) \tag{4.2}$$

In Eq. (4.2), the hypothesis of speech presence is denoted by $\mathcal{H}_1$. The current noise PSD estimate $\widehat{|N_k(\ell)|^2}$ is therefore a soft weighting between the noisy observation $|Y_k(\ell)|^2$ and the recursively smoothed noise PSD estimate $\widehat{\sigma}^2_{N,k}(\ell-1)$ from the previous time frame, where the weighting factor $P(\mathcal{H}_1|Y_k(\ell))$ is the SPP given $Y_k(\ell)$. The current noise PSD estimate in Eq. (4.2) is therefore updated only when speech was absent. To compute the estimate in Eq. (4.2), the SPP is derived (Gerkmann and Hendriks, 2012):

$$P(\mathcal{H}_1|Y_k(\ell)) = \left(1 + \frac{1-P(\mathcal{H}_1)}{P(\mathcal{H}_1)}\big(1+\xi_{\mathcal{H}_1}\big)\exp\left(-\frac{|Y_k(\ell)|^2}{\widehat{\sigma}^2_{N,k}(\ell-1)}\frac{\xi_{\mathcal{H}_1}}{\xi_{\mathcal{H}_1}+1}\right)\right)^{-1} \tag{4.3}$$

From Eq. (4.3), the prior probability of speech, $P(\mathcal{H}_1)$, and the a priori SNR expected under speech presence, $\xi_{\mathcal{H}_1}$, are required. If $P(\mathcal{H}_1) = 0.5$, it can be shown that $\xi_{\mathcal{H}_1} = 8\,\text{dB}$ (Gerkmann and Hendriks, 2012). The noise PSD estimate in Eq. (4.2) is then recursively smoothed over time using a time constant of

71.7 ms (Gerkmann and Hendriks, 2012) to obtain the $\widehat{\sigma}^2_{N,k}(\ell)$:

$$\widehat{\sigma}^2_{N,k}(\ell) = \alpha_{PSD}\widehat{\sigma}^2_{N,k}(\ell-1) + (1-\alpha_{PSD})\widehat{|N_k(\ell)|^2} \tag{4.4}$$

The estimates are combined into $M = 22$ non-overlapping auditory CI channels spaced between 244.7 Hz and 7279.2 Hz. Finally, the estimates are converted into the *a posteriori* SNR estimate, $\widehat{\gamma_k}(\ell)$, and then the *a priori* SNR estimate, $\widehat{\xi_k}(\ell)$:

$$\widehat{\gamma_k}(\ell) = \frac{|Y_k(\ell)|^2}{\widehat{\sigma}^2_{N,k}} \tag{4.5}$$

$$\widehat{\xi_k}(\ell) = \begin{cases} \bar{\gamma}(k,\ell) - 1, & \text{if } \bar{\gamma}(k,\ell) > 1 \\ 0, & \text{otherwise,} \end{cases} \tag{4.6}$$

4.2.2 The noise-reduction and channel-selection strategies

The estimated SNRs were utilized in both a noise-reduction strategy and in two channel-selection strategies (Fig. 4.1). In the noise-reduction strategy, called "NR-SPP&ACE", a set of gain values were computed from a Wiener gain function, which has been optimized for CI recipients (Mauger et al., 2012b). The Wiener gain function is described in *Appendix C.1*. The set of gain values was applied to the noisy envelopes in the CI signal path as a pre-processing step to ACE, and from the processed envelopes n-of-m channels with the largest processed amplitudes were selected for electrical stimulation. In the fixed channel-selection strategy based on the SNR criterion, called "CS-SNR", estimated SNRs were directly used to select the n-of-m channels with the highest SNRs. In the adaptive channel-selection strategy where the channel-selection is based on an estimated IBM, called "CS-IBM", an LC of 0 dB was first applied to the SNRs to determine which channels were speech-dominated and therefore candidates for stimulation. This LC has previously been used in Hu and Loizou (2008). In *Appendix C.2*, the impact of choosing different local criteria in the channel selection has been analyzed on electrodogram error rates. In order to keep the stimulation rate the same as in the CI recipients' everyday mapping, only up to n channels were then stimulated in each cycle, where n is the number of maxima selected for ACE in each recipients' default map.

The strategies were compared to ACE as the reference, and to ACE in combination with MS-based noise reduction (i.e., "NR-MS&ACE") within which the estimated SNRs were computed using MS (Martin, 2001).

Figure 4.2 depicts the true and estimated noise PSDs for a randomly-selected sample sentence from the Bamford-Kowal-Bench (BKB)-like corpus (Bench et al., 1979) mixed with multi-talker babble at 0 SNR (Fig. 4.2a) and at 5 dB SNR (Fig. 4.2b). The SPP-based estimator was able to track the changes in the true noise PSD faster than the MS-based estimator, as is evident by the fact that the MS-based estimator led to a larger underestimation of the true noise PSD than the SPP-based estimator in most time frames. To quantify the accuracy of the noise PSD, the logarithmic estimation error, LogErr, was adopted (Hendriks et al., 2008) across time frames, ℓ, and frequency channels, m:

$$\text{LogErr} = \frac{10}{LM} \sum_{l=1}^{L} \sum_{m=1}^{M} \log_{10} \frac{\sigma_{N,m}^2(\ell)}{\widehat{\sigma}_{N,m}^2(\ell)} \tag{4.7}$$

The logarithmic estimation error was computed for 10 sentences from a randomly-chosen list mixed with multi-talker babble at 0 dB and 5 dB SNRs for each of the two noise PSD estimators. The improvement of the SPP-based estimator was 1.1 dB relative to the MS-based estimator when averaged across sentences and SNRs. This result is consistent with data from Gerkmann and Hendriks (2012), where improvements of about 1 dB were found for similar conditions.

4.2.3 Study design

The subjects participated in two sessions, and in each session four different strategies were tested. In Session 1, the strategies ACE, "NR-MS&ACE", "CS-SNR" and "CS-IBM" were tested in speech-weighted noise to compare the channel-selection strategies with existing speech coding strategies, as well as to assess the impact of altering the SNR-based channel selection from fixed to adaptive. In Session 2, ACE, "NR-MS&ACE", "NR-SPP&ACE" and the best performing SNR-based channel-selection strategy of the two in Session 1 were tested in a more challenging multi-talker babble condition. Furthermore, Session 2 investigated if an improved noise PSD estimator accuracy in the context of noise-reduction

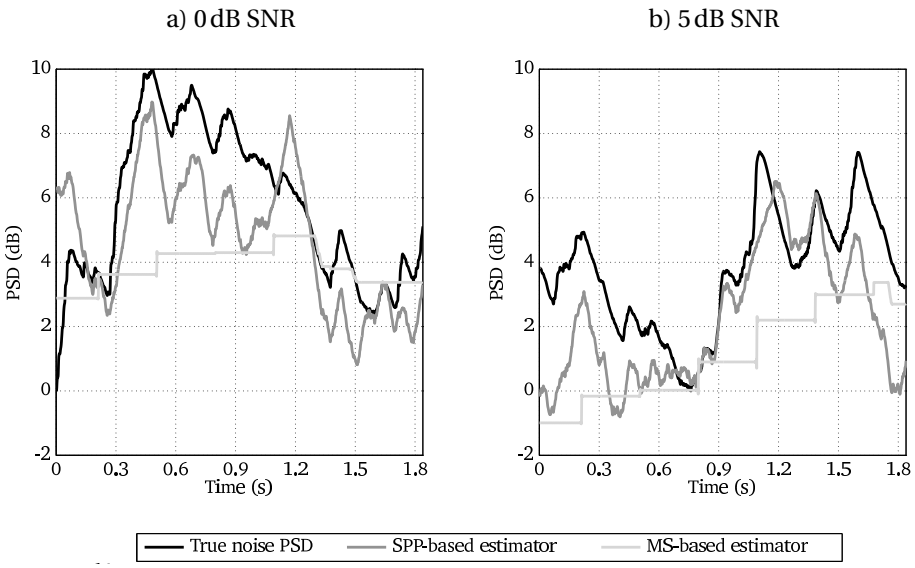


Figure 4.2: True and estimated noise PSDs for a noisy speech signal mixed with multi-talker babble at 0 SNR (Fig. 4.2a) and at 5 dB SNR (Fig. 4.2b). Noise PSDs are shown for the CI channel number 10 with a center frequency of 1406.9Hz. The true noise PSD was recursively smoothed over time using a first-order low-pass filter, similar to previous studies (Hendriks et al., 2008; Gerkmann and Hendriks, 2012). A smoothing time constant of 151.9 ms was used corresponding to $\alpha = 0.993$ with the current step size of 1 ms.

could translate into higher measured speech intelligibility.

4.2.4 Hardware and stimuli

The speech coding strategies were implemented with SIMULINK in a real-time system developed by Cochlear Limited. The system consisted of a host PC, and a target xPC which executed the SIMULINK models in real-time. The microphone input was recorded directly from the behind the ear sound processor worn by the recipient on the same ear as the implant. The real-time system processed the noisy speech signal, and the selected channels were stimulated by producing a radio frequency output through a coil that transmitted the stimulation sequence directly to the recipient's implant. In all strategies, automatic sensitivity control and adaptive dynamic range optimization were enabled (Mauger et al., 2014).

The BKB-like corpus consists of 80 lists with 16 sentences per list. The root mean square (RMS) levels of all individual sentences were equalized, and the Long Term Average Spectrum (LTAS) of the noise was adjusted to the LTAS of

speech (Byrne et al., 1994). The sentences were mixed with speech-weighted noise or multi-talker babble from 20-talkers, and subsequently presented at 0 degrees azimuth 1.2 m in front of the recipients at 65 dB sound pressure level (SPL) via a loudspeaker in a sound isolated booth.

4.2.5 Subjects

Twelve CI recipients participated in the study. Biographical data is presented in Table 4.1. The subject age spanned from 37 to 85 years with a median age of approximately 69 years. The CI usage time ranged from 1 to 13 years with a median of 8 years. All but one subject were stimulated with $n = 8$ maxima out of $m = 22$ electrodes while the remaining subject was stimulated with $n = 12$ maxima.

Table 4.1: Biographical data for the 12 CI subjects.

Subject	Age (yrs)	CI usage (yrs)	n maxima	m electrodes	Rate (pps)
S1	72	4	8	21	900
S2	67	5	8	22	900
S3	78	7	8	22	1200
S4	56	13	8	22	900
S5	77	8	8	22	1200
S6	85	8	8	22	900
S7	58	8	8	22	900
S8	80	8	12	22	900
S9	37	5	8	22	500
S10	72	9	8	22	900
S11	48	9	8	22	500
S12	65	1	8	22	900

4.2.6 Procedure

Subjects were tested with an adaptive speech reception threshold (SRT) task in noise and a quality ranking task in each session. The Australian Sentence Test in Noise (AUSTIN) (Dawson et al., 2013) was used to derive a single SRT from a psychometric curve fit to the percentage of correct morpheme scores for 24 BKB-like sentences from across two lists[1]. Each strategy was evaluated

[1]This calculation rule provides better reliability than the Hearing In Noise Test (HINT) calculation rule (Dawson et al., 2013).

with two runs. The strategies were evaluated in a repeated measures design, and the test order was counterbalanced within the session and randomized across subjects. Subjects were familiarized with the SRT test by presenting 16 processed sentences with the strategy “CS-IBM”.

Sound quality was rated with a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test. Samples from an audio book of Australian stories with the same speaker were presented at 65 dB SPL, and the speech was mixed with the noises at the SRT of ACE prior to the processing with the strategies in each session. In addition to the four strategies, a reference condition was added. The reference condition was created by lowering the noise level by 10 dB in the “ACE” strategy. The subjects were allowed to switch between the strategies instantaneously and as many times as necessary by using a touch screen. Two questions were asked in the MUSHRA, namely “How clear do you perceive the speech?” and “How annoying is the noise?”. The two questions were presented to the subjects two times, and the obtained ratings were averaged.

In addition to these two tasks, Session 2 also tested monosyllabic word recognition in quiet using the consonant–vowel nucleus–consonants (CNCs). See *Appendix C.3* for a description of the test and a discussion of the results.

4.2.7 Statistical analysis

A linear mixed effect model was constructed for each set of SRTs, quality ratings and the CNCs in each session. In all the models, strategies were considered a fixed factor and subjects a random effect. In addition, runs were considered a random effect for the set of SRTs. The fixed factor, the random effects and the interactions were initially included in the model. The model was then reduced by performing a backward elimination of all random interactions that were non-significant on a 5% significance level. Fixed factor levels were tested at a 5% significance level. To visualize the data, the predicted least-squares means and 95% confidence limits of the least-squares means were extracted from the model. To assess any difference between strategies, the differences of the least-squares means were computed in pairwise comparisons where the p values were adjusted following the Tukey multiple comparison testing.

4.3 Results

Figure 4.3 shows the measured SRTs in speech-weighted noise in Session 1 (Fig. 4.3a) and in multi-talker babble in Session 2 (Fig. 4.3b) with the four different strategies. Individual SRTs for each of the twelve CI recipients are shown with different symbols. In addition, horizontal black bars indicate the least square means, and the gray shaded boxes show the 95% confidence limits of the least square means predictions from the fitted linear mixed effect models.

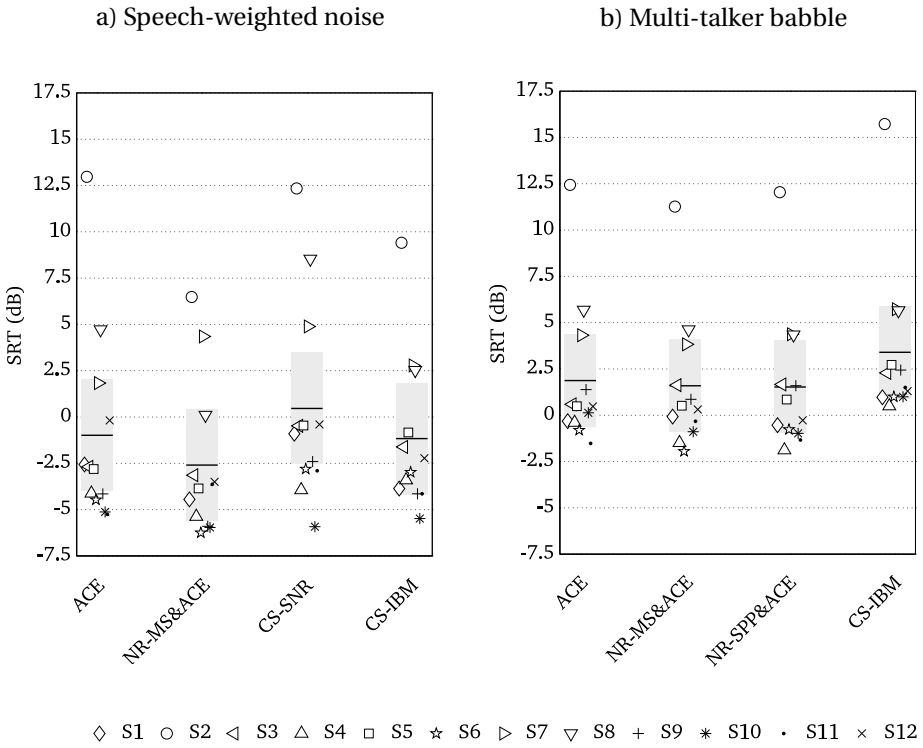


Figure 4.3: Measured SRTs in speech-weighted noise (Fig. 4.3a) and in multi-talker babble (Fig. 4.3b). Individual SRTs for each of the twelve CI recipients are shown with different symbols. The horizontal black bars illustrate the least square means, and the gray shaded boxes show the 95% confidence limits of the least square means predictions.

In both Figs. 4.3a and 4.3b, a high SRT variability was observed across the CI recipients. In the strategy “ACE”, three subjects (“S2”, “S7” and “S8”) showed a poor performance with up to 12 dB in SRT, while the remaining subjects performed very well with SRTs as low as −6 dB in speech-weighted noise and around 0 dB in the multi-talker babble condition.

Figure 4.4 shows corresponding quality ratings with the recipients in Ses-

sion 1 (Fig. 4.4a) and Session 2 (Fig. 4.4b). The reference program ("REF") was included in the comparison and the recipients should ideally rate this program with the highest score in the MUSHRA. Subjects "S2" and "S8" also had difficulties in rating the reference program the highest in Figs. 4.4a and 4.4b. Most likely, providing these subjects with a 10 dB SNR improvement in the reference program was not sufficient for them to hear a difference in clarity or noise annoyance.

4.3.1 Evaluation of the noise-reduction strategies

The "NR-MS&ACE" and "NR-SPP&ACE" strategies were first compared. In Fig. 4.3b, half of the subjects ("S1", "S4", "S8", "S10", "S11" and "S12") showed lower SRTs with the "NR-SPP&ACE" strategy than with the "NR-MS&ACE" strategy, but no statistically significant difference between the two strategies was observed. Thus, the results suggest that speech intelligibility does not improve significantly with the more accurate SPP-based noise PSD estimator relative to the MS-based estimator. In Fig. 4.4b, the speech clarity in both strategies was rated significantly higher over the "ACE" strategy. Furthermore, the noise annoyance was rated significantly lower in both strategies over the "ACE". The two strategies ("NR-SPP&ACE" and "NR-MS&ACE") did, however, not differ from each other statistically in terms of the sound quality measures.

In addition to these findings, the existing noise-reduction strategy ("NR-MS&ACE") improved the SRT compared to ACE alone by about 1.6 dB in speech-weighted noise (see Fig. 4.3a). This is consistent with previously-reported findings (Dawson et al., 2011; Mauger et al., 2012a,b), e.g. about 2 dB was obtained in Dawson et al. (2011). Finally, neither of the two noise-reduction strategies ("NR-MS&ACE" or "NR-SPP& ACE") improved speech intelligibility significantly relative to ACE in the multi-talker babble (Fig. 4.3b). In comparison, a small improvement of about 7% of the measured word recognition score (WRS) has been reported in this noise type (Mauger et al., 2012a).

4.3.2 Evaluation of the channel-selection strategies

The fixed ("CS-SNR") and the adaptive channel-selection ("CS-IBM") strategies were then compared (Fig. 4.3a). The "CS-IBM" strategy was found to decrease the mean SRT scores by 1.63 dB as compared to the "CS-SNR" strategy ($p < 0.01$).

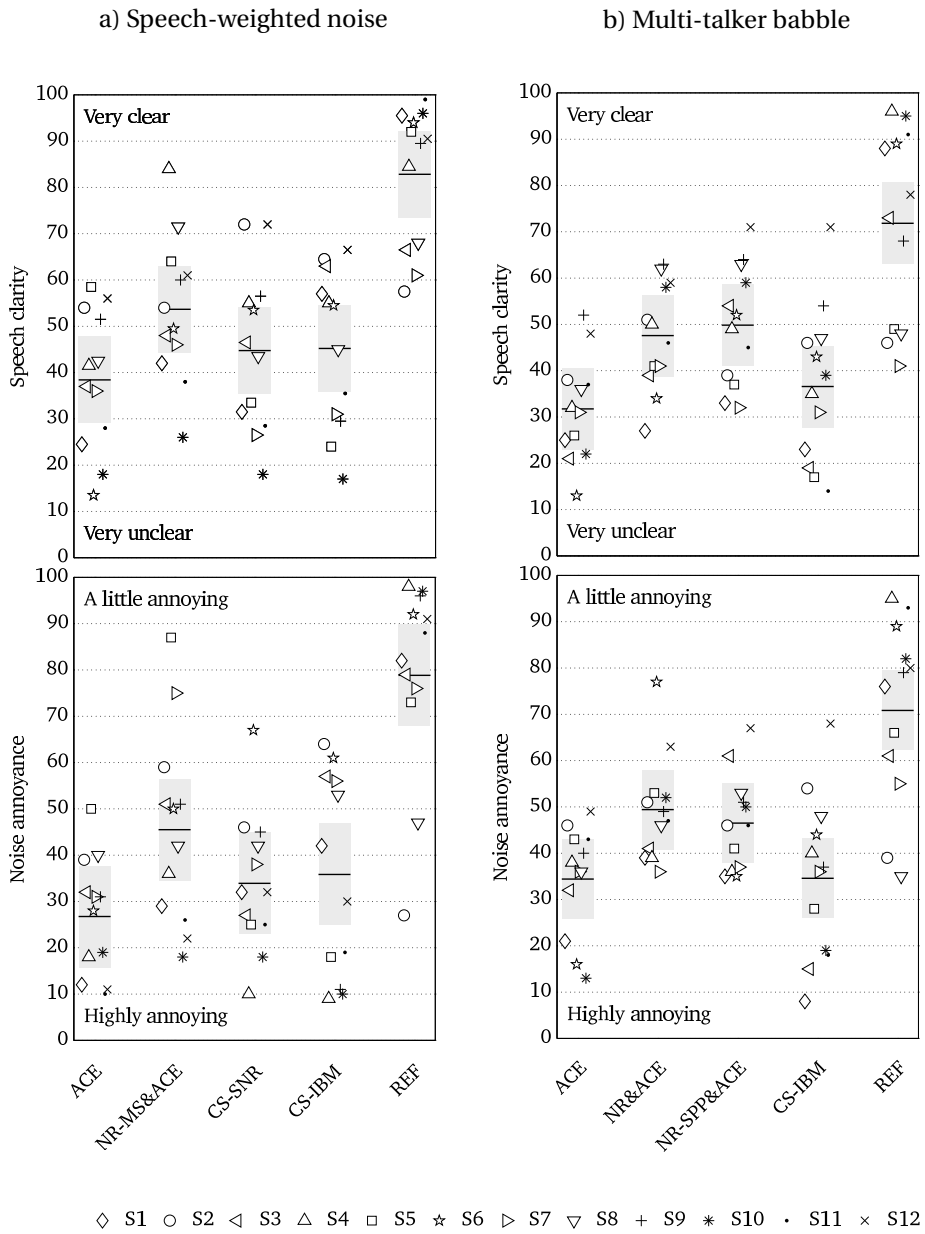


Figure 4.4: Measured quality ratings in speech-weighted noise in Session 1 (left panels) and in multi-talker babble in Session 2 (right panels). Individual ratings for each of the twelve CI recipients are shown with different symbols. The horizontal black bars illustrate the least square means, and the gray shaded boxes show the 95% confidence limits of the least square means predictions.

The adaptively changing channel-selection therefore improved the speech intelligibility relative to the fixed channel-selection in the CI recipients.

Comparing the "CS-IBM" strategy with the ACE strategy (Fig. 4.3a), there was no significant difference in mean SRTs and in sound quality ratings in speech-weighted noise (Fig. 4.3a), despite that almost half of the subjects ("S1", "S2", "S8", "S10" and "S12") had lower SRTs with the "CS-IBM" strategy and these same subjects also rated the speech clarity higher and the noise annoyance lower with the "CS-IBM" strategy over the ACE strategy (Fig. 4.4a). Moreover, the SRT actually increased by 1.53 dB ($p < 0.0001$), i.e. speech intelligibility was worse with the "CS-IBM" strategy in the presence of multi-talker babble (Fig. 4.3b). Therefore, neither of the two SNR-based channel-selection strategies improved speech intelligibility relative to ACE.

4.4 Discussion

4.4.1 Improved noise power estimation in noise-reduction strategies

The SPP-based noise PSD estimator, as proposed by Gerkmann and Hendriks (2012), was considered in the context of noise-reduction, instead of the MS-based estimator which is currently used in current CI processors (Dawson et al., 2011; Mauger et al., 2012a,b). The findings of the current study demonstrate that the SPP-based noise PSD estimator is more accurate in tracking the true noise PSD than the MS-based estimator in the multi-talker babble condition in terms of the logarithmic estimation error, which confirm previous findings (Gerkmann and Hendriks, 2012). Nevertheless, the results from the CI listener study showed that the improved accuracy in noise PSD estimation does not translate into an increase in measured speech intelligibility. Two points may help explain this observation. First, the SPP-based noise PSD estimate changed more rapidly over time and the gain values therefore also varied more quickly over time. The CI recipients are accustomed to a more slowly changing noise-reduction strategy in their everyday sound processors ("NR-MS&ACE"), since this noise-reduction strategy is most likely used on a daily basis and, in most cases, has been used for many years. A lack of familiarity with the SPP-based noise-reduction strategy may thus have affected the results. Secondly, the logarithmic estimation error, as described in Eq. (4.7), does not reveal in which time frames and frequency channels the noise PSD estimator is tracking the true noise PSD with high accuracy (i.e., whether the accuracy is high when speech is present or absent).

The results therefore suggest that the logarithmic estimation error is not a good predictor of the speech intelligibility outcome.

4.4.2 Analysis of the logarithmic estimation error

Previous studies (Qazi et al., 2013; Kressner et al., 2017) have employed segmentation methods to investigate the effects of noise-induced errors in speech coding. Specifically, Kressner et al. (2017) divided the ideal electrodogram into three temporal regions, based on the stimulation activity across CI frequency channels: speech gaps, the so-called speech transitions which define the boundaries between the speech segments and speech gaps[1] and the speech segments. In Qazi et al. (2013), findings suggested that CI recipients can tolerate significantly lower levels of noise in speech gaps and at the same time comparable levels in speech segments. In addition to these findings, Kressner et al. (2017) showed that the benefit of attenuating the noise-dominated channels in speech gaps was relatively small when the speech transitions in the ideal electrodogram at the same time had been degraded with errors.

In the present study, the logarithmic estimation error was used to assess the accuracy of the introduced SPP-based estimator against the existing MS-based estimator. To gain insights into *where* the introduced noise PSD estimator is more accurate, as compared to the existing estimator, the segmentation method of Kressner et al. (2017) is therefore applied. Figure 4.5 visualizes the segmentation of an ideal electrodogram generated from the sample sentence shown in Fig. 4.2 into the three temporal regions. In the lower panel of Fig. 4.5, the detected speech gaps, speech transitions and speech segments are marked. The estimated noise PSDs are then segmented accordingly, and the logarithmic estimation error is computed within each temporal region (Fig. 4.6). The logarithmic estimation error is lower within each of the temporal regions for the SPP estimator, as compared to the MS-based estimator, which indicate that it is more accurate in each of the regions. However, there is no significant difference in accuracy among the speech gaps, speech transitions and the speech segments with neither of the noise PSD estimators. In relation to the findings of Qazi et al.

[1] Qazi et al. (2013) only divided the regions into speech gaps and speech segments. However, Kressner et al. (2017) argued that it is necessary to also consider the speech transitions.

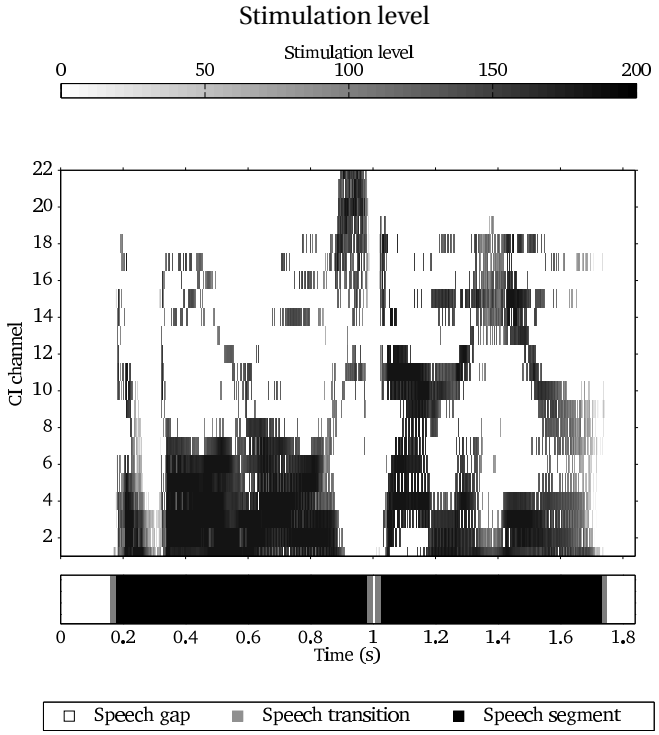


Figure 4.5: The ideal electrodogram (top panel) was obtained by processing the speech signal with ACE. Graytones correspond to a stimulation level between the T and C level. The ideal electrodogram was subsequently segmented into speech gaps (bottom panel), speech segments and speech transitions. The speech transition is marked with a gray color line and the offset of the transition by a dashed line. The applied segmentation method is described in Kressner et al. (2017).

(2013) and Kressner et al. (2017) and the present study, a very low logarithmic estimation error is most likely required in the speech gaps and speech transitions. Noise PSD estimation algorithms should therefore prioritize a low estimation error in both speech gaps and in speech transitions when employed in noise-reduction strategies in CIs. The segmentation of the logarithmic estimation error can be a useful tool for such an assessment.

4.4.3 Using noise power estimation in channel-selection strategies

Neither of the channel selection strategies improved the speech intelligibility relative to the well-established ACE strategy. There may be three possible explanations for this.

First and foremost, even though the SPP-based noise PSD estimator has

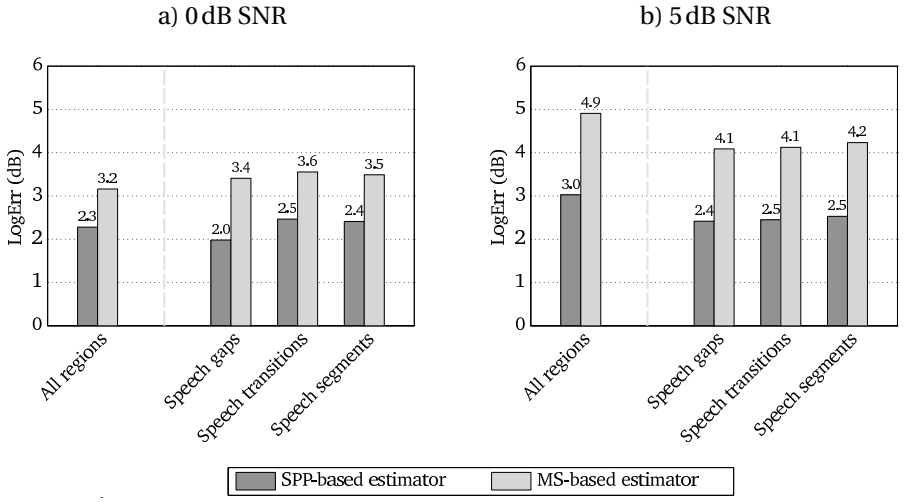


Figure 4.6: The logarithmic estimation error is computed by comparing the estimated noise PSD to the smoothed true noise PSD (Hendriks et al., 2008; Gerkmann and Hendriks, 2012). Errors are computed with the SPP-based and the MS-based noise PSD estimators for a multi-talker babble mixed with the previously-used sample sentence at 0 SNR (Fig. 4.6a) and at 5 dB SNR (Fig. 4.6b). The segmentation method, described in Kressner et al. (2017), is then employed. In Fig 4.5a, the ideal electrodogram has been segmented into three different temporal regions (speech gaps, speech transitions and speech segments). The error is then computed within each of the temporal regions.

decreased the logarithmic estimation error, it does not appear to be accurate enough for SNR-based channel selection, since performance with these strategies was not close to that obtained with SNR-based channel selection based on *a priori* SNRs (Hu and Loizou, 2008). To shed light on what is then a sufficient accuracy for SNR-based channel selection, the accuracy of the considered SNR estimation algorithm is compared with two previously-used algorithms (Hu et al., 2007; Hu and Loizou, 2010) in a post hoc analysis that follows the approach in Hu and Loizou (2008). Specifically, the ideal and estimated SNRs were compared separately to an LC to construct an IBM and an estimated IBM. A hit rate was then calculated as the percentage of correctly classified speech-dominated T-F units, and a false alarm rate was calculated as the percentage of incorrectly classified noise-dominated T-F units. The hit-false alarm (H-FA) rate was then used to measure the SNR estimation accuracy (Table 4.2). The estimation algorithm from Hu et al. (2007) is based on the noise PSD estimator of Cohen and Berdugo (2002), combined with modified SNR estimation of Ephraim and Malah (1984). The SNR estimation algorithm in the present study had a much higher hit rate and a lower false alarm rate

Table 4.2: Hit and false alarm rates for three SNR estimation algorithms in multi-talker babble at 5 dB SNR. Only hit and false alarm rates were available for this particular condition in previous studies (Hu et al., 2007; Hu and Loizou, 2010).

SNR estimation algorithm	Hit rate (%)	False alarm rate (%)	H-FA (%)
Hu et al. (2007)	53.19	17.41	35.78
Present study	71.21	15.22	55.99
Hu and Loizou (2010)	89.29	14.19	75.10

than the Hu et al. (2007) estimation algorithm. Therefore, it estimated the instantaneous SNR more accurately than the Hu et al. (2007) estimation algorithm[1]. However, the SNR estimation algorithm from the present study had an approximately 20% lower H - FA rate than the Hu and Loizou (2010) algorithm, which employed a speech segregation system with a high-complexity classifier to estimate the SNRs. These findings imply that the SNR estimation algorithm in the present study had a limited accuracy, as compared to the SNR estimation algorithm from Hu and Loizou (2010). Studies have shown that the H - FA rate has limitations in predicting speech intelligibility (Kressner et al., 2016; Bentsen et al., 2018b), and it is therefore important to emphasize that the H - FA rate can only be used to evaluate the SNR estimation accuracy. Whether the limited accuracy of the SNR estimation (compared to the Hu and Loizou (2010) algorithm) can explain the poor speech intelligibility remains speculative.

Secondly, a lack of training with the channel-selection strategies by the CI recipients may have influenced the performance. Prior to the testing, the CI recipients were only familiarized with the SNR-based channel-selection strategies by presenting 16 sentences whereas the existing speech coding strategies (ACE and "NR-MS&ACE') are both integrated in the participants' everyday sound processors.

Finally, an experimental constraint was that only up to n channels were stimulated in the adaptively-changing channel-selection strategy, where $n = 8$ for most of the participants. In comparison, up to 16 (out of the 16) channels were available for stimulation in Hu and Loizou (2008) when the SNR was high. However, this limited subset of n-of-m channels seems sufficient enough for

[1] The Hu et al. (2007) estimation algorithm was never applied for SNR-based channel selection in Hu and Loizou (2008), presumably because of the low SNR estimation accuracy.

ACE, such that it is unlikely to be the primary explanation for the lack of any speech intelligibility improvement.

4.4.4 From fixed to adaptively-changing channel selection

The impact of altering the SNR-based channel selection from fixed to adaptive was also investigated. Results indicated that the adaptively-changing channel selection resulted in a substantially higher speech intelligibility than the fixed channel selection in speech-weighted noise. In the adaptive channel-selection strategy, the LC was applied to force the channel selection to be adaptively changing between 0 and n channels across stimulation cycles. This is illustrated in Fig. 4.7. Figure 4.7a shows an ideal electrodogram ("Speech-in-quiet") and Figs. 4.7b-e show estimated electrodograms with the four different strategies from Session 1 for a comparison across strategies. In Figs. 4.7d-e, n channels were selected in both electrodograms in regions with speech only (compare to the ideal electrodogram in Fig. 4.7a). However, in the electrodogram generated using the adaptively-changing channel-selection strategy (Fig. 4.7e) fewer than n channels were stimulated in the CI recipients when the instantaneous SNR was low in the speech gaps, and therefore, the CI recipients were exposed to less noise-induced stimulation. Reducing stimulation in speech gaps has previously been shown to be important for improving speech intelligibility in noise, because CI recipients can tolerate significantly lower levels of noise in the speech gaps than in the speech segments (Qazi et al., 2013).

4.5 Conclusion

Speech intelligibility outcomes with ACE are limited in noisy conditions with low SNRs, since the channels with the largest amplitudes can be noise-dominated instead of speech-dominated, and since ACE always selects a fixed number of channels when the signal amplitude is above a predefined threshold, irrespective of speech presence or absence. A range of different speech-coding strategies have been developed to alleviate these shortcomings. Either noise-dominant channels can be attenuated prior to ACE with noise-reduction strategies or, alternatively, channels can be selected based on an SNR criterion. Both types of strategies, however, require an accurate noise PSD estimation stage. This study investigated the impact of noise PSD estimation in noise-reduction and channel-

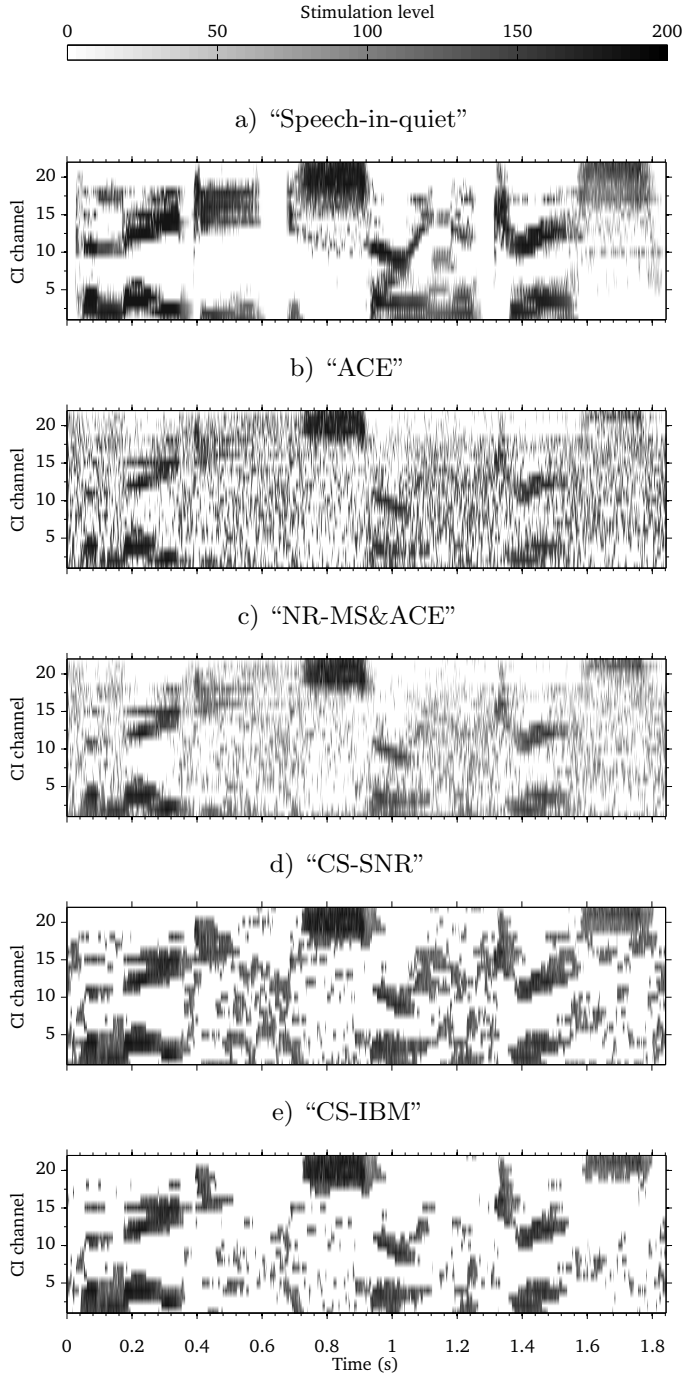


Figure 4.7: The ideal electrodogram (“Speech-in-quiet”) was obtained by processing the speech signal with ACE. In Figs. 4.7b-e, the speech signal was mixed with speech-weighted noise at 0 dB SNR and processed with the four strategies in Session 1. Graytones correspond to a stimulation level between the T and C level.

selection strategies. Overall, the results of the present study indicate that a noise PSD estimation with improved noise-tracking capabilities, and therefore a higher accuracy, does not necessarily translate to increased speech intelligibility when the noise PSD estimation is utilized for noise reduction nor for when it is utilized for channel selection. Noise PSD estimation algorithms which have a higher accuracy in both the speech gaps and in the speech transitions are required, and the potential of such algorithms for speech coding in CIs can be assessed with the introduced segmentation analysis. In addition, the impact of altering the SNR-based channel-selection criterion from fixed to adaptive was investigated in the present study. The LC is important in the SNR-based channel selection for reducing the noise-induced stimulation in the CI recipients.

Acknowledgments

This study was carried out in collaboration with Cochlear Limited, and clinical testing was conducted at Cochlear Melbourne, Australia. We thank Dr Kerrie Plant and Evelyn Do for audiological and clinical assistance.

5

General discussion

5.1 Summary and implications of the main findings

This thesis investigated three approaches within computational speech segregation based on ideal time-frequency mask estimation, and the approaches were evaluated in the framework of single-channel noise reduction in normal-hearing (NH) listeners and cochlear-implant (CI) recipients in various adverse conditions.

Chapter 2 considered a speech segregation system, which employed a Gaussian mixture model (GMM) classifier for each subband, to estimate the ideal binary mask (IBM). The study focus was on how to exploit contextual information in speech across time and frequency in computational speech segregation. Specifically, the impact of different strategies that exploit spectro-temporal context was investigated on measured speech intelligibility in NH listeners. Computing the so-called ”delta features” in the subband-based system, and appending them to the feature vector, led to higher measured speech intelligibility scores than employing a support vector machine-based integration stage after the GMM-based classification stage. Using the delta features was therefore found to be the better strategy to exploit context by the subband-based system. The delta features were subsequently explored in detail across subbands. The measured speech intelligibility scores increased with the amount of spectral information exploited, until reaching a plateau when seven subbands or more were included in the feature vector.

Chapter 3 considered a selection of components within recent and successful state-of-the-art approaches, based on deep neural networks (DNNs). The motivation of the study was to explore the roles and the relative contributions of the components by measuring speech intelligibility in NH listeners. The first component was the system architecture, which was changed

from a subband-based GMM classifier, used in Chapter 2, to a DNN in which mask units are predicted simultaneously across all subbands. A substantial improvement by 25.4 percentage points in measured speech intelligibility scores was obtained going from the subband-based GMM architecture to the DNN architecture. The second component was a widely-used time frame concatenation technique, which is often applied in DNN-based speech segregation systems to exploit the temporal context. However, the time frame concatenation technique did not lead to any significant improvement in measured speech intelligibility, presumable because of the relatively limited amount of training data was not sufficient to unfold the predictive power of the DNN. The final component was the ideal time-frequency mask which is often considered the learning objective in computational speech segregation systems. By changing the learning objective from the IBM to an ideal ratio mask (IRM), another improvement of 13.9 percentage points was achieved in terms of measured speech intelligibility scores. Thus, both components contribute to the success of the state-of-the-art approaches. By combining the two significant components, namely the DNN architecture and the IRM estimation, an intelligibility improvement of about $7-8$ percentage points, relative to noisy speech, was obtained in an adverse conditions in which a fluctuating six-talker noise degraded the speech at a low signal-to-noise ratio (SNR).

Findings presented in Chapters 2 and 3 have implications for the system design within computational speech segregation. First, that spectral context is important to be considered. By employing the DNN architecture in recent approaches, the spectral context is exploited in a broadband manner. This is not the case with subband-based approaches. Specifically, when using the delta features in the subband GMM-based system to exploit context, more than seven subbands did not increase the measured speech intelligibility any further. The subband GMM classifier was therefore limited in the capability to exploit the correlation across frequency of the feature space. On the other hand, the DNN was capable of exploiting spectral context more effectively than the corresponding subband GMM classifier. Presumably, because DNNs can handle higher-dimensional training data where an arbitrary number of input feature vectors can be mapped to an arbitrary number of outputs. These findings imply that the inherent ability of the DNN architecture to exploit

spectral context is effective, which makes an DNN desirable from a system design perspective. While findings in the current thesis emphasize that spectral context is important to exploit in computational speech segregation systems, it is unclear how much temporal context is required to increase the speech intelligibility outcome in listeners. Specifically, Chapter 3 demonstrated that the applied time frame concatenation had no effect on measured speech intelligibility. This result was rather surprising, but should be seen in light of two points. First, a small amount of training data was fed to the DNN-based system which can explain why the predictive power of the DNN was not unfolded. Secondly, "only" five past feature frames were appended to the current frame, resulting in an exploited temporal context of 40 ms. In other studies with feed-forward network architectures (Chen et al., 2016a; Healy et al., 2017), a larger number of past time frames and also future time frames have been concatenated, resulting in a much larger amount of temporal context of up to 200 ms. Recently, Chen and Wang (2017) has employed a recurrent neural network with long short term memory (LSTM), instead of the feed-forward network. They demonstrated that this recurrent network performs better than the feed-forward network in terms of intelligibility predictions. This network did not use future time frames, and it was argued, based on the predictions, that the ability of LSTM to capture long-term speech context is important (Chen and Wang, 2017). The impact of these time frame concatenation techniques has not been investigated on measured speech intelligibility, which is important to draw any decisive conclusions. A final implication of the findings involves the learning objective in computational speech segregation. Findings in Chapter 3 demonstrated a substantial perceptual advantage with the IRM as a learning objective, instead of the IBM, since estimated IRM values were able to convey more speech information (i.e., the mask values were more tolerant to misses). Although the binary classification makes the IBM a simpler benchmark to estimate, the IRM should be preferred as the learning objective when speech segregation systems are designed for noise-reduction applications.

Finally, *Chapter 4* considered an application of the estimated ideal time-frequency mask in speech-coding strategies in real-time CI processing. Specifically, the study investigated the impact of state-of-the-art noise power spectral density (PSD) estimation in a range of different speech-coding strategies to improve the speech intelligibility outcome, as compared to the advanced

combination encoder (ACE). Either noise-dominant channels were attenuated prior to the ACE with noise-reduction strategies or, alternatively, channels were selected based on SNRs, similar to how the IBM is constructed. The results in Chapter 4 indicate that a noise PSD estimation with improved noise-tracking capabilities, and therefore a higher accuracy, does not necessarily translate to increased speech intelligibility when the noise PSD estimation is utilized for noise reduction nor for when it is utilized for channel selection. A segmentation analysis indicated that a much higher accuracy in both the speech gaps and in the speech transitions are required. In addition, the impact of altering the SNR-based channel-selection criterion from a fixed to adaptively-changing across time was investigated. Specially, when the local criterion (LC) was applied to the estimated SNRs, an increase in measured speech intelligibility outcome was achieved. This adaptively-changing channel selection is therefore important in the SNR-based channel selection for reducing the noise-induced stimulation in the CI recipients. Overall, the findings imply that novel speech-coding strategies should employ estimation algorithms which have a much higher accuracy in both the speech gaps and in the speech transitions, and where the number of stimulated channels is changed adaptively over time to reduce the exposure to noise-induced stimulation in CI recipients.

5.2 Improving the generalization ability to unseen conditions

In speech segregation systems, it is important to consider the ability to generalize to acoustic conditions which are not seen during training (i.e., "mismatches"). These mismatches can include noise segments, SNRs, noise types, speakers and the signal shaping (e.g., from mobile phones or due to room reverberation).

In Chapters 2 and 3, different segments of a fluctuating six-talker noise were considered between system training and testing. Specifically in Chapter 2, the measured speech intelligibility scores decreased quite substantially when unseen noise segments were considered during testing (Fig. 2.5 in Sec. 2.4.2), as compared to conditions in which the same short noise segments were used during training and testing (Fig. 2.3a in Sec. 2.4.1). By appending more subbands to the feature vector, the feature space increased in size

and a larger amount of spectral information was revealed. Appending more subbands led to an improved ability of the subband GMM-based system to generalize to unseen noise segments. However, the subband GMM classifier was limited in the capability to exploit the increased feature space for more than seven subbands. Overall, the subband GMM-based system therefore demonstrated a rather limited generalization ability to the considered mismatch. A moderate-complexity classifier of 16 GMMs was used when addressing the generalization ability of the system. *Appendix A* considered the system performance with 16 GMMs across the duration of the noise recording, from which the noise segments were randomly selected during training and testing. A stable system performance was found, when noise segments from a 50 s noise recording (and beyond) were randomly selected. On the other hand, a stable system performance was not obtained in May and Dau (2014b) when using a higher number of GMMs in the speech segregation system (e.g., 64, 128 or 256). Therefore, the subband GMM-based system would most likely have worsened the generalization ability if a high-complexity subband GMM classifier had been selected instead. GMMs are generative and probabilistic models in which a set of Gaussian components are used to model the feature space. A shortcoming of these models, as demonstrated here, is the limited capacity which can affect the generalization ability to unseen conditions.

Large-scale training of speech segregation systems is an important component in computational speech segregation to handle mismatched conditions, but was not employed in the speech segregation systems presented in this thesis. In Chen et al. (2016a), a speech segregation system demonstrated an ability to generalize to a range of novel noise types and SNRs during system testing. In Kolbæk et al. (2017), mismatches in SNR, noise type and speaker identity were each handled successfully by three DNN-based systems. These three systems were SNR-dependent, noise-type-dependent and speaker-identity-dependent, respectively. What these recent approaches have in common is a feed-forward network architecture more complex than the feed-forward network architecture, presented in Chapter 3. The network architecture typically contains multiple hidden layers with hundreds of nodes in each layers. Better non-linear activation functions, such as rectified linear units, have been included, and dropout and maxout have been used during the network training (Chen et al., 2016a; Healy et al., 2017; Kolbæk et al., 2017). Moreover, the DNNs have been trained

with a large amount of data which consist of multiple conditions of different noise segments, SNRs, noise types or speakers. The capability of the DNNs to scale in size, and therefore handle a large training data set, is most likely key to the improved generalization ability. In Chapter 3, the considered DNN-based system was able to generalize to unseen noise segments of the fluctuating six-talker noise; however, the speech intelligibility improvement over noisy was small. It is worth noting that the study goal in Chapter 3 was to study the roles of other components than the large-scale training. Retraining the DNN-based system with a larger data set would most likely have improved the generalization ability to the specific mismatch of noise segments. Furthermore, by choosing a more complex network architecture and expanding the training session with multiple conditions, the speech segregation system may have been able to handle other mismatches, than considered in this thesis.

5.3 One cost function that correlates with measured speech intelligibility

Objective measures have previously been used to optimize the performance of computational speech segregation systems during the development stage. In this thesis, several discrepancies were observed between predictions of the objective measures and measured speech intelligibility. In Chapter 2, the measured speech intelligibility scores were compared to the extended short-term objective intelligibility (ESTOI) index and the H - FA rate[1]. A finding was that the ESTOI and the H - FA could not alone account for all of the measured observations. In *Appendix B*, the ESTOI predicted speech intelligibility improvements when no improvements were actually measured in the study presented in Chapter 3. In Chapter 4, a decrease in the logarithmic estimation error did not translate to an increase in measured speech intelligibility within existing speech-coding strategies. Another objective measure, namely the electrodogram error rate (Mauger et al., 2012a; Hersbach, 2014) was computed based on electrodograms generated from a number of speech-coding strategies in *Appendix C.2*. Electrodogram error rates were not able to correctly predict the ranking of the speech coding strategies in the listener study, presented in Chapter 4. Thus, the findings across

[1]The H - FA rate was calculated as the difference between the percentage of correctly classified speech-dominated time-frequency (T-F) units (hit rate, H) and the percentage of incorrectly classified noise-dominated T-F units (false alarm rate, FA) (Kim et al., 2009).

all thesis chapters emphasize the need for a single objective measure to correctly predict the speech intelligibility in listeners. Such an objective measure is highly relevant as a cost function to assess and optimize the design of computational speech segregation systems with noise-reduction applications.

5.4 Perspectives for future studies

Significant progress has been made over the past years in computational speech segregation in the context of single-channel noise reduction, and speech intelligibility improvements, relative to the noisy speech, have been demonstrated in normal-hearing and hearing-impaired listeners. Today's speech segregation systems utilize advanced machine-learning techniques, and they are able to generalize to unseen noise segments, SNRs, noise types as well as speakers. Even in a "two-talker" condition, i.e. a speech signal in the presence of a single competing-talker, speech intelligibility improvements have been demonstrated (Healy et al., 2017). Until now, the speech segregation systems have been trained to handle specific mismatches (i.e., being noise-independent, SNR-independent etc.). However, a system able to improving speech intelligibility in listeners, trained independent of SNRs, noise types or speakers, is still considered the "The Holy Grail". Kolbæk et al. (2017) trained a system on a number of SNRs, noise types as well as speakers, and evaluated the system in a condition which considered an unseen noise type and speaker simultaneously; however, this system failed to improve speech intelligibility. One direction for future work is therefore to investigate if such an "independent system" can be constructed. The network architecture within computational speech segregation has been improved over the last couple of years, and will most likely continue to evolve in the next couple of years. Instead of a feed-forward network, a recurrent neural network with LSTM can be employed, since these networks are powerful for time series predictions (Chen and Wang, 2017). In addition, and as highlighted in Sec. 5.2, large-scale training is important. The data set for training should include as many variations as possible which can be obtained by using multiple conditions of SNRs, noise types and speakers. In the feature extraction stage, different features can be considered which capture as many relevant characteristics of the speech and noise as possible. In the present thesis, only the amplitude modulation spectrogram (AMS) features (Kollmeier and Koch, 1994; Tchorz and Kollmeier,

2003) have been extracted which capture the modulations of the speech and noise. Several other auditory-inspired features can be appended as well, e.g. pitch-based features or Gammatone log energy-based features. A systematic feature study can help assessing the potential of the features in the framework of large-scale training. Measuring speech intelligibility should be part of the evaluation.

Another direction for future work can consider practical applications in hearing aids or CIs. The state-of-the-art noise PSD estimator of Gerkmann and Hendriks (2012) is generic and real-time applicable with a low latency, which means it is feasible on a digital signal processing (DSP) chip for hearing aids or CIs. However, the SNR estimation accuracy was insufficient to improve the speech intelligibility outcome as observed in Chapter 4. Currently, it is unclear if a noise PSD estimator can be constructed which tracks the true noise PSD more accurately in particular in the speech gaps and speech transitions. DSP technology evolves and neural network processors with low-power consumption are already available for embedded system applications. In the future, it may be possible to process speech segregation systems with a low number of network weights on DSPs in hearing aids or CIs[1]. Low-latency values are required in hearing aids and CIs which will constraint the allowed processing time of certain blocks (e.g., the feature extraction). If a speech segregation system with low complexity can fulfil some of the fundamental generalization requirements, it may be feasible in such a practical application. The system shall be robust to temporal changes in the noise (i.e., be able to handle mismatches of noise segments and SNRs in a certain range). In addition, the system shall generalize to a range of unseen noise types, since the user may encounter a range of different noise types on a daily basis. In Goehring et al. (2017), a low-complexity and low-latency DNN-based system[2] was trained to estimate a set of gains for noise reduction as pre-processing in ACE. Significant speech intelligibility improvements were obtained with this approach in CI recipients, relative to ACE. The system is able to generalize to unseen noise segments, SNRs and to a novel speaker in two out of three of

[1]The system training is usually time consuming but can be done offline on graphical processing units, and therefore latency and computational cost are not crucial aspects for network training.

[2]With an almost similar network architecture to the one considered in Chapter 3.

the considered noise types. Therefore, it satisfies some of the fundamental generalization requirements. However, the system does not generalize to all unseen noise types yet. Such a system is highly relevant to estimate SNRs in the channel-selection strategies for speech coding described in Chapter 4. While speech segregation systems with low complexity may be embedded in hearing aids or CIs, an "independent" system will most likely be computationally too complex and therefore require too many network weights. Such a system may be embedded on a smartphone connected to the hearing aid or CI[1].

Finally, several discrepancies have been demonstrated with the objective measures considered in this thesis. Therefore, yet another direction for future work is the development of a speech intelligibility prediction model capable of predicting the different conditions presented in this thesis[2]. Such a speech intelligibility prediction model would be highly relevant in numerous research labs that focus on the research and the development of computational speech segregation systems.

[1] However, audio streaming between the smartphone and the device may be a challenge if the system requires high-latency processing

[2] A substantial amount of measured speech intelligibility data from listeners studies has become available for model testing.

Bibliography

Bench, J., Å. Kowal, and J. Bamford (1979). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children". *British journal of audiology* 13.3, pp. 108–112.

Ephraim, Y. and D. Malah (1984). "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator". *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.6, pp. 1109–1121.

Ephraim, Y. and D. Malah (1985). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator". *IEEE Trans. Audio, Speech, Signal Process.* 33.2, pp. 443–445.

Wilson, B. et al. (1988). "Comparative studies of speech processing strategies for cochlear implants". *The Laryngoscope* 98.10, pp. 1069–1077.

Zelinski, R. (1988). "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms". In: *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE, pp. 2578–2581.

Festen, J. M. and R. Plomp (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing". *J. Acoust. Soc. Amer.* 88.4, pp. 1725–1736.

McDermott, H. J., C. M. Mckay, and A. E. Vandali (1992). "A new portable sound processor for the University of Melbourne/Nucleus Limited multielectrode cochlear implant". *The Journal of the Acoustical Society of America* 91.6, pp. 3367–3371.

Byrne, D. et al. (1994). "An international comparison of long-term average speech spectra". *J. Acoust. Soc. Amer.* 96.4, pp. 2108–2120.

Kollmeier, B. and R. Koch (1994). "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction". *J. Acoust. Soc. Amer.* 95.3, pp. 1593–1602.

Lyon, R. (1997). "All-pole models of auditory filtering". *Diversity in auditory mechanics*, pp. 205–211.

Cooke, M., P. Green, L. Josifovski, and A. Vizinho (2001). "Robust automatic speech recognition with missing and unreliable acoustic data". *Speech Commun.* 34.3, pp. 267–285.

Dillon, H. (2001). *Hearing aids*. Vol. 362. Boomerang press Sydney.

Dreschler, W. A., H. Verschuure, C. Ludvigsen, and S. Westermann (2001). "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment". *Audiology* 40.3, pp. 148–157.

Martin, R. (2001). "Noise power spectral density estimation based on optimal smoothing and minimum statistics". *IEEE Trans. Speech Audio Process.* 9.5, pp. 504–512.

Simmer, K. U., J. Bitzer, and C. Marro (2001). "Post-filtering techniques". In: *Microphone arrays*. Springer, pp. 39–60.

Cohen, I. and B. Berdugo (2002). "Noise estimation by minima controlled recursive averaging for robust speech enhancement". *IEEE Signal Process. Lett.* 9.1, pp. 12–15.

Tchorz, J. and B. Kollmeier (2003). "SNR estimation based on amplitude modulation analysis with applications to noise suppression". *IEEE Trans. Audio, Speech, Lang. Process.* 11.3, pp. 184–192.

Loizou, P. C., A. Lobo, and Y. Hu (2005). "Subspace algorithms for noise reduction in cochlear implants". *J. Acoust. Soc. Amer.* 118.5, pp. 2791–2793.

Wang, D. (2005). "On ideal binary mask as the computational goal of auditory scene analysis". In: *Speech separation by humans and machines*. Ed. by P. Divenyi. Springer, pp. 181–197.

Brungart, D. S., P. S. Chang, B. D. Simpson, and D. Wang (2006). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation". *J. Acoust. Soc. Amer.* 120.6, pp. 4007–4018.

Cooke, M. (2006). "A glimpsing model of speech perception in noise". *J. Acoust. Soc. Amer.* 119.3, pp. 1562–1573.

Srinivasan, S., N. Roman, and D. Wang (2006). "Binary and ratio time-frequency masks for robust speech recognition". *Speech Commun.* 48.11, pp. 1486–1501.

Barker, J. and M. Cooke (2007). "Modelling speaker intelligibility in noise". *Speech Commun.* 49.5, pp. 402–417.

Hu, Y. and P. C. Loizou (2007). "A comparative intelligibility study of single-microphone noise reduction algorithms". *J. Acoust. Soc. Amer.* 122.3, pp. 1777–1786.

Hu, Y., P. C. Loizou, N. Li, and K. Kasturi (2007). "Use of a sigmoidal-shaped function for noise attenuation in cochlear implants". *J. Acoust. Soc. Amer.* 122.4, EL128–EL134.

Bentler, R., Y.-H. Wu, J. Kettel, and R. Hurtig (2008). "Digital noise reduction: Outcomes from laboratory and field studies". *Int. J. Audiol.* 47.8, pp. 447–460.

Gannot, S. and I. Cohen (2008). "Adaptive beamforming and postfiltering". In: *Springer handbook of speech processing*. Springer, pp. 945–978.

Hendriks, R. C., J. Jensen, and R. Heusdens (2008). "Noise tracking using DFT domain subspace decompositions". *IEEE Trans. Audio, Speech, Lang. Process.* 16.3, pp. 541–553.

Hu, Y. and P. C. Loizou (2008). "A new sound coding strategy for suppressing noise in cochlear implants". *J. Acoust. Soc. Amer.* 124.1, pp. 498–509.

Li, N. and P. C. Loizou (2008). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction". *J. Acoust. Soc. Amer.* 123.3, pp. 1673–1682.

Wang, D., U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner (2008). "Speech perception of noise with binary gains". *J. Acoust. Soc. Amer.* 124.4, pp. 2303–2307.

Kim, G., Y. Lu, Y. Hu, and P. C. Loizou (2009). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners". *J. Acoust. Soc. Amer.* 126.3, pp. 1486–1494.

Kjems, U., J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang (2009). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech". *J. Acoust. Soc. Amer.* 126.3, pp. 1415–1426.

Nielsen, J. B. and T. Dau (2009). "Development of a Danish speech intelligibility test". *Int. J. Audiol.* 48.10, pp. 729–741.

Hendriks, R. C., R. Heusdens, and J. Jensen (2010). "MMSE based noise PSD tracking with low complexity". In: *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, pp. 4266–4269.

Hu, Y. and P. C. Loizou (2010). "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users". *J. Acoust. Soc. Amer.* 127.6, pp. 3689–3695.

Chang, C.-C. and C.-J. Lin (2011). "LIBSVM: A library for support vector machines". *ACM TIST* 2 (3). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 27:1–27:27.

Dawson, P. W., S. J. Mauger, and A. A. Hersbach (2011). "Clinical evaluation of signal-to-noise ratio–based noise reduction in Nucleus® cochlear implant recipients". *Ear Hear.* 32.3, pp. 382–390.

Loizou, P. C. and G. Kim (2011). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions". *IEEE transactions on audio, speech, and language processing* 19.1, pp. 47–56.

Nielsen, J. B. and T. Dau (2011). "The Danish hearing in noise test". *Int. J. Audiol.* 50.3, pp. 202–208.

Taal, C. H., R. C. Hendriks, R. Heusdens, and J. Jensen (2011). "An algorithm for intelligibility prediction of time–frequency weighted noisy speech". *IEEE Trans. Audio, Speech, Lang. Process.* 19.7, pp. 2125–2136.

Brons, I., R. Houben, and W. A. Dreschler (2012). "Perceptual effects of noise reduction by time-frequency masking of noisy speech". *J. Acoust. Soc. Amer.* 132.4, pp. 2690–2699.

Gerkmann, T. and R. C. Hendriks (2012). "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay". *IEEE Trans. Audio, Speech, Lang. Process.* 20.4, pp. 1383–1393.

Han, K. and D. L. Wang (2012). "A classification based approach to speech segregation". *J. Acoust. Soc. Amer.* 132.5, pp. 3475–3483.

Jensen, J. and R. C. Hendriks (2012). "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions". *IEEE Transactions on Audio, Speech, and Language Processing* 20.1, pp. 92–102.

Mauger, S. J., K. Arora, and P. W. Dawson (2012a). "Cochlear implant optimized noise reduction". *J. Neural Eng.* 9.6, p. 065007.

Mauger, S. J., P. W. Dawson, and A. A. Hersbach (2012b). "Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction". *J. Acoust. Soc. Amer.* 131.1, pp. 327–336.

May, T., S. van de Par, and A. Kohlrausch (2012a). "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation". *IEEE Trans. Audio, Speech, Lang. Process.* 20.7, pp. 2016–2030.

May, T., S. van de Par, and A. Kohlrausch (2012b). "Noise-robust speaker recognition combining missing data techniques and universal background modeling". *IEEE Trans. Audio, Speech, Lang. Process.* 20.1, pp. 108–121.

Dawson, P. W., A. A. Hersbach, and B. A. Swanson (2013). "An adaptive Australian sentence test in noise (AuSTIN)". *Ear Hear.* 34.5, pp. 592–600.

Hazrati, O. and P. C. Loizou (2013). “Comparison of two channel selection criteria for noise suppression in cochlear implants”. *J. Acoust. Soc. Amer.* 133.3, pp. 1615–1624.

Healy, E. W., S. E. Yoho, Y. Wang, and D. L. Wang (2013). “An algorithm to improve speech recognition in noise for hearing-impaired listeners”. *J. Acoust. Soc. Amer.* 134.6, pp. 3029–3038.

May, T. and T. Dau (2013). “Environment-aware ideal binary mask estimation using monaural cues”. In: *Proc. WASPAA*. New Paltz, NY, USA, pp. 1–4.

Narayanan, A. and D. Wang (2013). “Ideal ratio mask estimation using deep neural networks for robust speech recognition”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, pp. 7092–7096.

Qazi Obaid ur Rehman van Dijk, B., M. Moonen, and J. Wouters (2013). “Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility”. *Hearing research* 299, pp. 79–87.

Wang, Y. and D. Wang (2013). “Towards scaling up classification-based speech separation”. *IEEE Trans. Audio, Speech, Lang. Process.* 21.7, pp. 1381–1390.

Hersbach, A. A. (2014). “Noise reduction for cochlear implants”. Not publicly accessible. PhD thesis. Electrical and Electronic Engineering, The University of Melbourne.

Hummersone, C., T. Stokes, and T. Brookes (2014). “On the ideal ratio mask as the goal of computational auditory scene analysis”. In: *Blind Source Separation*. Springer, pp. 349–368.

Mauger, S. J., C. D. Warren, M. R. Knight, M. Goorevich, and E. Nel (2014). “Clinical evaluation of the Nucleus® 6 cochlear implant system: Performance improvements with SmartSound iQ”. *International journal of audiology* 53.8, pp. 564–576.

May, T. and T. Dau (2014a). “Computational speech segregation based on an auditory-inspired modulation analysis”. *J. Acoust. Soc. Amer.* 136.6, pp. 3350–3359.

May, T. and T. Gerkmann (2014). “Generalization of supervised learning for binary mask estimation”. In: *Proc. IWAENC*. Juan les Pins, France, pp. 154–187.

May, T. and T. Dau (2014b). “Requirements for the evaluation of computational speech segregation systems”. *J. Acoust. Soc. Amer.* 136.6, EL398–EL404.

Wang, Y., A. Narayanan, and D. Wang (2014). "On training targets for supervised speech separation". *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 22.12, pp. 1849–1858.

Chen, J., Y. Wang, and D. Wang (2015). "Noise perturbation improves supervised speech separation". In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer, pp. 83–90.

Healy, E. W., S. E. Yoho, J. Chen, Y. Wang, and D. Wang (2015). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type". *J. Acoust. Soc. Amer.* 138.3, pp. 1660–1669.

Jensen, J. and M. S. Pedersen (2015). "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications". In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, pp. 5728–5732.

Kressner, A. A. and C. J. Rozell (2015). "Structure in time-frequency binary masking errors and its impact on speech intelligibility". *J. Acoust. Soc. Amer.* 137.4, pp. 2025–2035.

May, T., T. Bentsen, and T. Dau (2015). "The role of temporal resolution in modulation-based speech segregation". In: *Proc. Interspeech*. Dresden, Germany, pp. 170–174.

Bentsen, T., T. May, A. A. Kressner, and T. Dau (2016). "Comparing the influence of spectro-temporal integration in computational speech segregation". In: *Proc. Interspeech*. San Francisco, USA, pp. 170–174.

Chen, J., Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy (2016a). "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises". *J. Acoust. Soc. Amer.* 139.5, pp. 2604–2612.

Chen, J., Y. Wang, and D. Wang (2016b). "Noise perturbation for supervised speech separation". *Speech Commun.* 78, pp. 1–10.

Jensen, J. and C. H. Taal (2016). "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers". *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 24.11, pp. 2009–2022.

Kressner, A. A. and C. J. Rozell (2016). "Cochlear implant speech intelligibility outcomes with structured and unstructured binary mask errors". *J. Acoust. Soc. Amer.* 139.2, pp. 800–810.

Kressner, A. A., A. Westermann, J. M. Buchholz, and C. J. Rozell (2016). "Cochlear implant speech intelligibility outcomes with structured and unstructured binary mask errors". *J. Acoust. Soc. Amer.* 139.2, pp. 800–810.

Zhang, X.-L. and D. Wang (2016). “A deep ensemble learning method for monaural speech separation”. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 24.5, pp. 967–977.

Chen, J. and D. Wang (2017). “Long short-term memory for speaker generalization in supervised speech separation”. *J. Acoust. Soc. Amer.* 141.6, pp. 4705–4714.

Gelderblom, F. B., T. V. Tronstad, and E. M. Viggen (2017). “Subjective intelligibility of deep neural network-based speech enhancement”. In: *Proc. Interspeech*. Stockholm, Sweden, pp. 1968–1972.

Goehring, T., F. Bolner, J. J. Monaghan, B. van Dijk, A. Zarowski, and S. Bleeck (2017). “Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users”. *Hearing research* 344, pp. 183–194.

Healy, E. W., M. Delfarah, J. L. Vasko, B. L. Carter, and D. Wang (2017). “An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker”. *J. Acoust. Soc. Amer.* 141.6, pp. 4230–4239.

Kolbæk, M., Z.-H. Tan, and J. Jensen (2017). “Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems”. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 25.1, pp. 153–167.

Kressner, A. A., T. May, R. M. Thaarup Høegh, A. K. Juhl, T. Bentsen, and T. Dau (2017). “Investigating the effects of noise-estimation errors in simulated cochlear implant speech intelligibility”. *Proceedings of the International Symposium on Auditory and Audiological Research 6*.

Bentsen, T., T. May, A. A. Kressner, and T. Dau (2018a). “The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility”. *PLOS ONE* 13.5.

Bentsen, T., A. A. Kressner, T. Dau, and T. May (2018b). “The impact of exploiting spectro-temporal context in computational speech segregation”. *J. Acoust. Soc. Amer.* 143.1, pp. 248–259.

Contributions to Hearing Research

Vol. 1: *Gilles Pigasse*, Deriving cochlear delays in humans using otoacoustic emissions and auditory evoked potentials, 2008.

Vol. 2: *Olaf Strelcyk*, Peripheral auditory processing and speech reception in impaired hearing, 2009.

Vol. 3: *Eric R. Thompson*, Characterizing binaural processing of amplitude-modulated sounds, 2009.

Vol. 4: *Tobias Piechowiak*, Spectro-temporal analysis of complex sounds in the human auditory system, 2009.

Vol. 5: *Jens Bo Nielsen*, Assessment of speech intelligibility in background noise and reverberation, 2009.

Vol. 6: *Helen Connor*, Hearing aid amplification at soft input levels, 2010.

Vol. 7: *Morten Løve Jepsen*, Modeling auditory processing and speech perception in hearing-impaired listeners, 2010.

Vol. 8: *Sarah Verhulst*, Characterizing and modeling dynamic processes in the cochlea using otoacoustic emissions, 2010.

Vol. 9: *Sylvain Favrot*, A loudspeaker-based room auralization system for auditory research, 2010.

Vol. 10: *Sébastien Santurette*, Neural coding and perception of pitch in the normal and impaired human auditory system, 2011.

Vol. 11: *Iris Arweiler*, Processing of spatial sounds in the impaired auditory system, 2011.

Vol. 12: *Filip Munch Rønne*, Modeling auditory evoked potentials to complex stimuli, 2012.

Vol. 13: *Claus Forup Corlin Jespersgaard*, Listening in adverse conditions: Masking release and effects of hearing loss, 2012.

Vol. 14: *Rémi Decorsière*, Spectrogram inversion and potential applications for hearing research, 2013.

Vol. 15: *Søren Jørgensen*, Modeling speech intelligibility based on the signal-to-noise envelope power ratio, 2014.

Vol. 16: *Kasper Eskelund*, Electrophysiological assessment of audiovisual integration in speech perception, 2014.

Vol. 17: *Simon Krogholt Christiansen*, The role of temporal coherence in auditory stream segregation, 2014.

Vol. 18: *Márton Marschall*, Capturing and reproducing realistic acoustic scenes for hearing research, 2014.

Vol. 19: *Jasmina Catic*, Human sound externalization in reverberant environments, 2014.

Vol. 20: *Michał Fereczkowski*, Design and evaluation of individualized hearing-aid signal processing and fitting, 2015.

Vol. 21: *Alexandre Chabot-Leclerc*, Computational modeling of speech intelligibility in adverse conditions, 2015.

Vol. 22: *Federica Bianchi*, Complex-tone pitch representations in the human auditory system , 2016.

Vol. 23: *Johannes Zaar*, Measures and computational models of microscopic speech perception, 2016.

Vol. 24: *Gusztáv Lőcsei*, Lateralized speech perception with normal and impaired hearing, 2016.

Vol. 25: *Johannes Käsbach*, Characterizing apparent source width perception, 2016.

Vol. 26: *Suyash Narendra Joshi*, Modelling auditory nerve responses to electrical stimulation, 2017.

Vol. 27: *Henrik Gert Hassager*, Characterizing perceptual externalization in listeners with normal, impaired and aided-impaired hearing, 2017.

Vol. 28: *Richard Ian McWalter*, Perceptual and neural response to sound texture, 2017.

Vol. 29: *Jens Cubick*, Investigating distance perception, externalization and speech intelligibility in complex acoustic environments, 2017.

Vol. 30: *Gerard Encina Llamas*, Characterizing cochlear hearing impairment using advanced electrophysiological methods, 2017.

Vol. 31: *Christoph Scheidiger*, Assessing speech intelligibility in hearing-impaired listeners, 2018.

Vol. 32: *Alan Wiinberg*, Perceptual effects of non-linear hearing aid amplification strategies, 2018.

Vol. 33: *Thomas Bentsen*, Computational speech segregation inspired by principles of auditory processing, 2018.

A

Comparing the influence of spectro-temporal integration in computational speech segregation[a]

Abstract

The goal of computational speech segregation systems is to automatically segregate a target speaker from interfering maskers. Typically, these systems include a feature extraction stage in the front-end and a classification stage in the back-end. A spectro-temporal integration strategy can be applied in either the front-end, using the so-called delta features, or in the back-end, using a second classifier that exploits the posterior probability of speech from the first classifier across a spectro-temporal window. This study systematically analyzes the influence of such stages on segregation performance, the error distributions and intelligibility predictions. Results indicated that it could be problematic to exploit context in the back-end, even though such a spectro-temporal integration stage improves the segregation performance. Also, the results emphasized the potential need of a single metric that comprehensively predicts computational segregation performance and correlates well with intelligibility. The outcome of this study could help to identify the most effective spectro-temporal integration strategy for computational segregation systems.

A.1 Introduction

Computational speech segregation systems attempt to automatically segregate a target signal from interfering noise. One frequently-used approach is to

[a] This chapter is based on: Bentsen, T., T. May, A. A. Kressner, and T. Dau (2016). Comparing the influence of spectro-temporal integration in computational speech segregation. In: Proc. Interspeech. San Francisco, USA, pp. 170–174. 17th Annual Conference of the International Speech Communication Association, San Francisco, USA.

construct an ideal binary mask (IBM) by retaining only those time-frequency (T-F) units that are target-dominated (Wang, 2005). Many studies have used the IBM to segregate a target speech signal from a noisy mixture and demonstrated large intelligibility improvements (Brungart et al., 2006; Wang et al., 2008; Kjems et al., 2009). However, *a priori* knowledge about the target and interferer is rarely available in realistic conditions and therefore, the goal of computational speech segregation systems is to obtain an estimated binary mask (EBM) given the noisy speech.

Despite high levels of interfering noise, speech-dominated T-F units tend to cluster in spectro-temporal regions, forming so-called *glimpses*, and the size of these glimpses has been shown to correlate well with speech intelligibility scores from normal-hearing listeners (Cooke, 2006). Consequently, several studies have tried to explore spectro-temporal context in computational segregation systems. One strategy is to exploit context in the front-end by using so-called delta features (Kim et al., 2009), which capture feature variations across time and frequency at the expense of a higher dimensional feature vector. Alternatively, spectro-temporal context can be exploited in the classification back-end by employing a two-layer segregation stage (Healy et al., 2013; May and Dau, 2014a). Specifically, the posterior probability of speech presence obtained from a first classifier is learned by a second classifier across a spectro-temporal window, where the amount of integration can be controlled by the size of the window function (May and Dau, 2014a).

To date, the effectiveness of computational segregation systems and the benefit of spectro-temporal integration strategies have been primarily evaluated using a technical metric, namely the H - FA, which quantifies segregation performance by calculating the difference between the percentage of correctly classified speech-dominated T-F units (hit rate, H) and the percentage of incorrectly classified noise-dominated T-F units (false alarm rate, FA) (Kim et al., 2009; Han and Wang, 2012; Healy et al., 2013; May and Dau, 2013, 2014a; May and Dau, 2014b). However, there is evidence suggesting that speech intelligibility scores are highly dependent on the distribution of mask errors rather than the overall H - FA rate (Kressner and Rozell, 2015), and this questions the applicability of the H - FA as the sole metric to optimize or evaluate computational segregation systems. The clustering

of the speech-dominated T-F units in glimpses suggests that a certain type of structure is inherently embedded in the IBM. However, depending on the choice of the spectro-temporal integration strategy in either the front-end or the back-end, it might have different consequences on the error distribution in the EBM.

The goal of the present study is, therefore, to systematically analyze the influence of spectro-temporal integration strategies in the front-end and the back-end of a speech segregation system using not only the H - FA, but also by considering the distribution of errors and the impact on predicted speech intelligibility using the short-term objective intelligibility (STOI) metric (Taal et al., 2011). In previous studies (Kim et al., 2009; Healy et al., 2013), the same short noise recording has been used for training and testing. In such experimental setups, a classification-based segregation system can then potentially capture all characteristics of the signals (May and Dau, 2014b). A second goal is, therefore, to analyze the potential influence of the noise duration on each of the spectro-temporal integration strategies.

A.2 The speech segregation system

The segregation system consisted of a feature extraction front-end and a classification back-end (May et al., 2015), as shown in Fig. A.1. The target signal was reconstructed by applying the EBM to the subband signals of the noisy speech, as illustrated by the dashed line. Each processing stage is described in detail in the following.

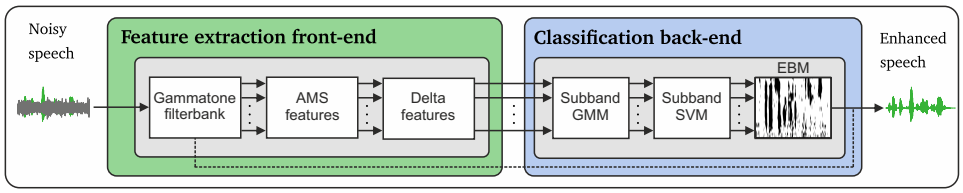


Figure A.1: Block diagram of the segregation system that shows the main blocks of the feature extraction front-end and the classification back-end. The dashed line illustrates the reconstruction of the target by applying the EBM to the subband signals of the noisy speech.

A.2.1 Feature extraction front-end

The distinct characteristics of speech and noise components were captured by amplitude modulation spectrogram (AMS) features (Tchorz and Kollmeier, 2003; Kim et al., 2009; May and Dau, 2014a; May et al., 2015). To derive these, the noisy speech was sampled at a rate of 16 kHz and decomposed into 31 frequency channels by a Gammatone filterbank, whose center frequencies were equally spaced on the equivalent rectangular bandwidth (ERB) scale between 80 and 7642 Hz. The envelope in each subband was extracted by half-wave rectification and low-pass filtering with a cutoff frequency of 1 kHz. Then, each envelope was normalized by its median that was computed over the entire signal, which was shown to improve the generalization to unseen acoustic conditions (e.g., signal-to-noise ratios (SNRs) and room reverberation) (May and Dau, 2014a; May and Gerkmann, 2014). The normalized envelopes were then processed by a modulation filterbank that consisted of one first-order low-pass and five band-pass filters with logarithmically spaced center frequencies and a constant Q-factor of 1. The root mean square (RMS) value of each modulation filter was then calculated across time frames corresponding to 32 ms with 75 % overlap, resulting in a 6-dimensional feature vector for each T-F unit $\mathbf{A}(t,f) = \{M_1(t,f), \ldots, M_6(t,f)\}^T$.

Context was explored in the front-end by appending delta features across time (Δ_T) and frequency (Δ_F) (Kim et al., 2009; Han and Wang, 2012; May and Dau, 2013). The final feature vector for each individual T-F unit at time frame t and frequency channel f consisted of $\mathbf{X}(t,f) = \left[\mathbf{A}(t,f), \Delta_T\mathbf{A}(t,f), \Delta_F\mathbf{A}(t,f)\right]$, where:

$$\Delta_T\mathbf{A}(t,f) = \begin{cases} \mathbf{A}(2,f) - \mathbf{A}(1,f), & \text{if } t = 1 \\ \mathbf{A}(t,f) - \mathbf{A}(t-1,f), & \text{otherwise,} \end{cases} \tag{A.1}$$

$$\Delta_F\mathbf{A}(t,f) = \begin{cases} \mathbf{A}(t,2) - \mathbf{A}(t,1), & \text{if } f = 1 \\ \mathbf{A}(t,f) - \mathbf{A}(t,f-1), & \text{otherwise.} \end{cases} \tag{A.2}$$

The size of the feature vector including delta features then increased from 6 dimensions to 18 dimensions.

A.2.2 Classification back-end

The classification back-end consisted of a two-layer segregation stage (May and Dau, 2014a; May et al., 2015). In the first layer, a Gaussian mixture model (GMM) classifier was trained to represent the speech and noise-dominated AMS feature distributions ($\lambda_{1,f}$ and $\lambda_{0,f}$) for each subband f. To separate the feature vector into speech- and noise-dominated T-F units, a local criterion (LC) was applied to the *a priori* SNR. The GMM classifier output was given as the posterior probability of speech and noise $P\big(\lambda_{1,f}|\mathbf{X}\big(t,f\big)\big)$ and $P\big(\lambda_{0,f}|\mathbf{X}\big(t,f\big)\big)$, respectively. The second layer consisted of a linear support vector machine (SVM) classifier (Chang and Lin, 2011), which considered the posterior probability of speech $P\big(\lambda_{1,f}|\mathbf{X}\big(t,f\big)\big)$ across a spectro-temporal integration window $\mathcal{W}$ for each subband (May and Dau, 2014a):

$$\bar{\mathbf{X}}\big(t,f\big) := \big\{P\big(\lambda_{1,u}|\mathbf{X}(u,v)\big) : (u,v) \in \mathcal{W}\big(t,f\big)\big\}. \tag{A.3}$$

According to (May and Dau, 2014a), a causal and plus-shaped window function $\mathcal{W}$ was used here, whereas the window size with respect to time and frequency was controlled by Δt and Δf, respectively.

A.3 Evaluation

A.3.1 Stimuli

The speech material was taken from the Danish Conversational Language Understanding Evaluation (CLUE) database (Nielsen and Dau, 2009), which consists of 70 sentences for training and 180 sentences for testing. Noisy speech mixtures with an average duration of 2 s were created by mixing individual sentences with a stationary (ICRA1) and a fluctuating 6-talker (ICRA7) noise masker (Dreschler et al., 2001). Both maskers had the same Long Term Average Spectrum (LTAS) as the CLUE corpus. A randomly-selected noise segment was used for each sentence and the noise segment started 250 ms before the speech onset and ended 250 ms after the speech offset.

A.3.2 Model training

The segregation system was trained for each of the two noise maskers. To investigate the influence of the noise duration, different models were trained

with noise files that were limited to 5, 10, 50 s or the total duration of the noise recording (60 s for ICRA1 and 600 s for ICRA7). The first layer of the classification back-end consisted of a GMM classifier with 16 Gaussian components and diagonal covariance matrices. The GMM classifier was trained with the 70 training sentences that were mixed three times with a randomly-selected noise segment at −5, 0 and 5 dB SNR. The subsequent SVM classifier was trained with only 10 sentences mixed at −5, 0 and 5 dB SNR. Afterwards, a re-thresholding procedure was applied (Han and Wang, 2012; May and Dau, 2014a) using a validation set of 10 sentences. Both classifiers employed a LC of −5 dB.

A.3.3 Model evaluation

The segregation system was evaluated with 180 CLUE sentences that were not used during training. Each sentence was mixed with ICRA1 and ICRA7 noises at −5 and 0 dB SNR. To study the influence of the noise duration, the trained models were evaluated with the same noise recordings used during training. Similar to the training, the noise recordings were limited in duration to 5, 10, 50 s or the total duration of the noise recording. In addition, a different noise recording of the same noise type was used to test the ability of the segregation system to generalize to unseen noise fluctuations of the same kind.

Three different metrics were used for evaluation, namely the H - FA, the clustering parameter γ and the STOI metric. The clustering parameter γ was estimated by the graphical model described in (Kressner and Rozell, 2015). Given a binary mask, the graphical model predicts the amount of clustering γ as a single number, where $\gamma = 1.0$ reflects a mask with uniformly and randomly connected T-F units. Larger values (e.g., $\gamma = 2.0$) reflect binary masks with T-F units that are twice as likely to be in the same state as its neighboring units (Kressner and Rozell, 2015). The STOI measure is based on a short-term correlation analysis between the clean and the degraded speech (Taal et al., 2011) mapped to a value between 0 and 1. In the current study, STOI improvements (Δ STOI) were reported as the relative difference between the predicted STOI values for the processed and the unprocessed noisy speech signal.

A.3.4 Experimental setup

To systemically analyze the influence of spectro-temporal integration in the front-end and the back-end, the following four segregation models were tested, as listed in Tab. A.1. "No integration" denotes the model with no delta features in the front-end and no spectro-temporal integration in the back-end ($\Delta t = 1, \Delta f = 1$). "Front-end" includes the delta features. "Back-end" does not utilize delta features, but applies spectro-temporal integration in the back-end ($\Delta t = 3, \Delta f = 9$). "Front- & back-end" exploits both delta features in the front-end and spectro-temporal integration in the back-end ($\Delta t = 3, \Delta f = 9$).

A.4 Results

The performance of the four segregation models and the IBM is presented in Fig. A.2 as a function of the noise duration for the two noise maskers ICRA1 (left panels) and ICRA7 (right panels). The three different panels on each side show the H - FA rate (top panels), the clustering parameter γ (middle panels) and the STOI metric (lower panels) averaged across 180 sentences and two SNRs (−5 and 0 dB).

In general, the segregation models produced higher H - FA rates in the presence of the stationary ICRA1 noise than for the ICRA7 noise, presumably because it was more difficult to separate the speech modulations from the non-stationary 6-talker babble noise. For both noise maskers, the lowest H - FA rates were observed for the "No integration" model and the highest H - FA rates for "Front- & back-end". Also, larger H - FA rates were obtained for the "Back-end" than the "Front-end" model. Each spectro-temporal integration strategy has previously been shown to improve H - FA rates separately (Kim

Table A.1: Configurations of the speech segregation system.

Model	Front-end		Back-end	
	Delta features	Feature dimension	$\mathcal{W}$ size Δt	Δf
No integration	no	6	1	1
Front-end	yes	18	1	1
Back-end	no	6	3	9
Front- & back-end	yes	18	3	9

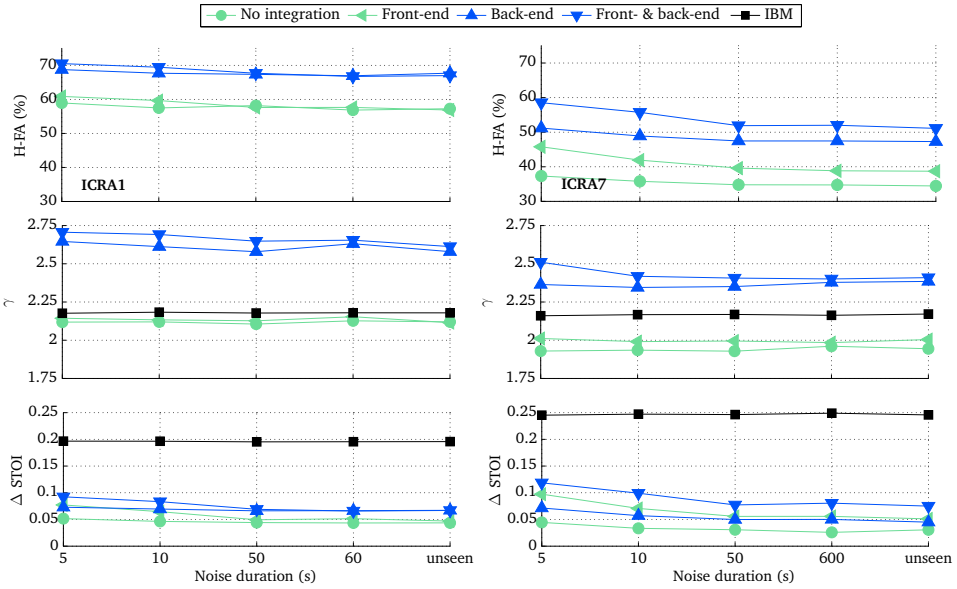


Figure A.2: H-FA, γ and STOI improvements for the four models and the IBM averaged across 180 sentences and SNRs (−5 and 0 dB) for ICRA1 (left panels) and ICRA7 (right panels). Average STOI values of the unprocessed noisy speech were 0.66 (ICRA1) and 0.63 (ICRA7).

et al., 2009; May and Dau, 2013, 2014a; May and Dau, 2014b). These previous results can be confirmed here for the ICRA7 noise by comparing both the "Back-end" and "Front-end" models with the "No integration" model.

The middle panels reveal that the IBM itself contains a certain amount of structure, presumably due to the compact representation of speech-dominated T-F units forming glimpses of the target signal. Also, reported values of γ from the model "No integration" are consistent with previous results (Kressner and Rozell, 2015; Kressner and Rozell, 2016). Most importantly, the γ values from models that exploited spectro-temporal context through the SVM classifier in the back-end (models "Back-end" and "Front- & back-end") are consistently larger than those from models where the SVM classifier did not incorporate contextual information across adjacent T-F units (models "No integration" and "Front-end"). On the contrary, the delta features alone do not seem to increase the amount of clustering in the mask.

In the bottom panels, the STOI improvement of the IBM indicates the largest possible intelligibility improvement that the segregation models can

achieve. The model "Front-end" produced larger STOI improvements than "Back-end" for the ICRA7 noise. Overall, the largest improvements were predicted for the model "Front- & back-end". In general, STOI predicted larger intelligibility improvements for ICRA7 than ICRA1.

Furthermore, Fig. A.2 demonstrates that the segregation system can capture all relevant signal characteristics when the same noise recording was used for training and testing, resulting in high H - FA rates and large STOI improvements for short noise durations. This trend was more pronounced for the non-stationary ICRA7 noise and decreased with longer noise duration. However, a moderate classifier complexity was chosen here (16 Gaussian components with diagonal covariance matrices), which was shown to reduce the risk of over-fitting the segregation system (May and Dau, 2014b). As a result, the generalization ability was improved, indicated by a stable system performance in terms of H - FA rates and STOI improvements for noise durations of 50 s and beyond. In contrast to the H - FA rates and STOI, the γ values stayed almost constant across the noise duration range.

Figure A.3 illustrates binary masks for one particular CLUE sentence mixed with ICRA7 noise at −5 dB SNR. Panel a) shows the IBM and panels b)-e) present the EBMs for the four tested models. The misclassified T-F units (misses and false alarms) are shown on top of the binary masks for a visualization of the error distributions. In addition, the evaluation metrics are shown in parenthesis. The effect of exploiting contextual knowledge in the back-end can be observed here. The panels d)-e) show masks with a larger amount of T-F clustering than the masks in panels b)-c). Obviously, the erroneous T-F units also become more structured.

A.5 Discussion and conclusion

Using the SVM classifier to exploit contextual knowledge in the back-end increased the H - FA rates but, at the same time, the amount of clustering (γ) in the masks was increased. In addition, the panels b)-e) in Fig. A.3 revealed that the increased amount of clustering also led to an increased clustering of the two types of mask errors (miss and false alarm). Previously, it has been argued that clustering of the two types of errors reduces the intelligibility

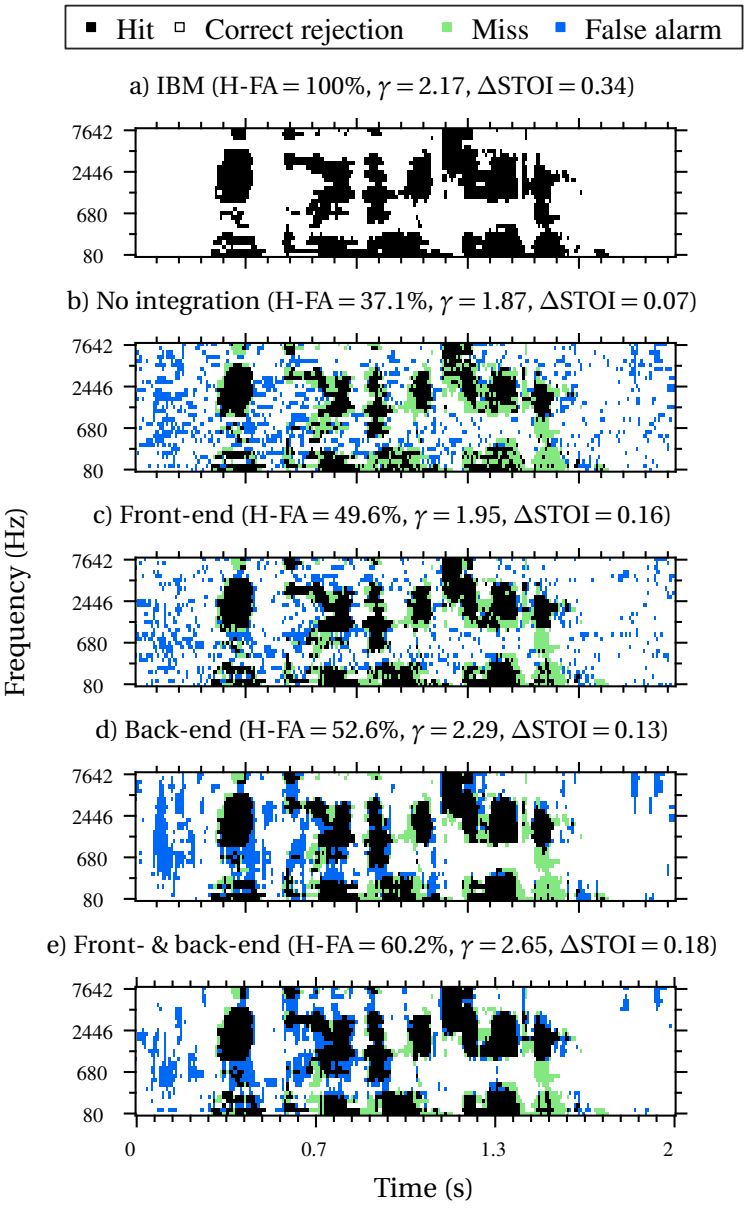


Figure A.3: Binary masks for a CLUE sentence mixed with ICRA7 noise at −5 dB SNR. Misses (target-dominated T-F units erroneously labeled as masker-dominated) and false alarms (masker-dominated T-F units erroneously labeled as target-dominated) are shown on top of the masks.

scores in comparison to the randomly distributed errors (Kressner and Rozell, 2015). This is supported by the predictions of the intelligibility scores with STOI, where larger improvements using the delta features than exploiting contextual knowledge in the back-end alone are predicted for the ICRA7

noise. This also means that, for an increased γ, a higher H - FA rate is required to obtain the same intelligibility score. It therefore seems problematic to exploit context in the back-end using a SVM classifier, even though such a spectro-temporal integration stage improves the H - FA rate (Healy et al., 2013; May and Dau, 2014a). The findings also suggest that using delta features might be a better spectro-temporal integration strategy in computational segregation systems, despite the fact that the H - FA rate does not increase as much as when exploiting contextual knowledge through a SVM classifier. However, it is necessary to confirm these findings with actual listening experiments.

In this study, both matched and unseen noise segments of the same noise type were used to evaluate classification-based segregation systems. As the ranking of the four models did not change with increasing noise durations, the findings of the influence of the spectro-temporal integration stage apply to both restricted and more realistic experimental setups with unseen noise segments of the same noise type. Future research will analyze the generalization ability of the segregation system to unseen noise types and will consider large-scale training (Chen et al., 2016a).

A recent study highlighted potential limitations of STOI in predicting the intelligibility of binary-masked speech (Kressner et al., 2016). Two observations from this study support these findings. Firstly, a higher H - FA rate does not necessarily lead to a larger STOI improvement as seen by comparing the “Front-end” and “Back-end” models. Secondly, if the SVM-based integration strategy in the back-end indeed has a detrimental effect on the intelligibility scores, it would imply that STOI over-predicts the model “Front- & back-end”. Thus, STOI alone would not account for all of the model differences described in this study. It emphasizes the potential need of a single metric that comprehensively predicts computational segregation performance and correlates well with intelligibility.

B

Comparing predicted and measured speech intelligibility in Bentsen et al. (2018a)[a]

The short-term objective intelligibility (STOI) index (Taal et al., 2011) and the extended short-term objective intelligibility (ESTOI) index (Jensen and Taal, 2016) are based on speech intelligibility prediction models commonly used to optimize the performance of computational speech segregation systems during the development stage (Wang et al., 2014; Zhang and Wang, 2016). The ESTOI is particularly used for modulated noise maskers. In this appendix, predictions from the ESTOI are compared to measured word recognition scores (WRSs) in *Chapter 3*. Table B.1 shows the ESTOI increase relative to noisy speech (ΔESTOI) and the WRS increase, also relative to noisy speech (ΔWRS). The six system configurations (described in Sec. 3.2.4) were evaluated using ICRA7 at −5 dB signal-to-noise ratio (SNR) and ESTOI values were averaged across all 180 test sentences.

Table B.1: ΔESTOI and ΔWRS relative to noisy speech with the six system configurations, described in Sec. 3.2.4. The system configurations were evaluated using ICRA7 at −5 dB SNR and ESTOI values were averaged across all 180 test sentences. WRS improvements are derived from the Paired Student's t-tests.

System configuration	ΔESTOI	ΔWRS (%)
GMM (IBM)	0.04	−31.1
GMM (IBM, 7 subbands)	0.08	−12.2
DNN (IBM)	0.11	−5.7
DNN (IBM, 40 ms)	0.11	−10.7
DNN (IRM)	0.14	8.2
DNN (IRM, 40 ms)	0.15	6.8

The predicted ΔESTOI values were larger for the deep neural network (DNN)-

[a]This appendix contains supplementary material for Chapter 3.

based system than the subband Gaussian mixture model (GMM)-based system. This is consistent with the ΔWRS values of the last column in Table B.1. Therefore, ESTOI correctly predicted the increase in measured speech intelligibility scores going from the subband GMM-based system to the DNN-based system. However, all predicted ΔESTOI values were positive in Table B.1 which indicated speech intelligibility improvements relative to noisy speech. This was not consistent with the ΔWRS values in Table B.1 where only improvements were observed for the "DNN (IRM)" and the "DNN (IRM, 40 ms)" system configurations. This particular finding highlights the discrepancy between predicted values of ESTOI and measured speech intelligibility scores.

C

Optimized Wiener gain function, electrodogram error rates, and an evaluation of the strategies in quiet [a]

C.1 A Wiener gain function optimized for CI recipients

In the "NR-SPP&ACE" strategy, the estimated signal-to-noise ratios (SNRs) are used to compute a set of gain values from a Wiener gain function, which has been found from a research study to be optimized for cochlear-implant (CI) recipients (Mauger et al., 2012b):

$$G = \left(\frac{\widehat{\xi_k}(\ell)}{\widehat{\xi_k}(\ell) + \alpha} \right)^{\beta} \tag{C.1}$$

Here, α and β denotes the gain threshold and gain slope, respectively. For the "NR-SPP&ACE" strategy, a gain threshold of 3 dB and a gain slope of 0.8 dB were used.

C.2 Strategy performance predictions using partial errors

Prior to testing with CI recipients, an error analysis was conducted with *partial error rates* (Hersbach, 2014). An ideal electrodogram was constructed by processing speech in quiet with "ACE" at 65 dB sound pressure level (SPL). Partial error components were computed by taking into account the stimulus intensity level in the electrodogram. Specifically, the difference in intensity level between the estimated and the ideal electrodogram was computed. A

[a] This appendix contains supplementary material for Chapter 4. The chapter is based on research in collaboration with Cochlear Limited (Dr Stefan Mauger) during an external research stay at Cochlear Melbourne, Australia.

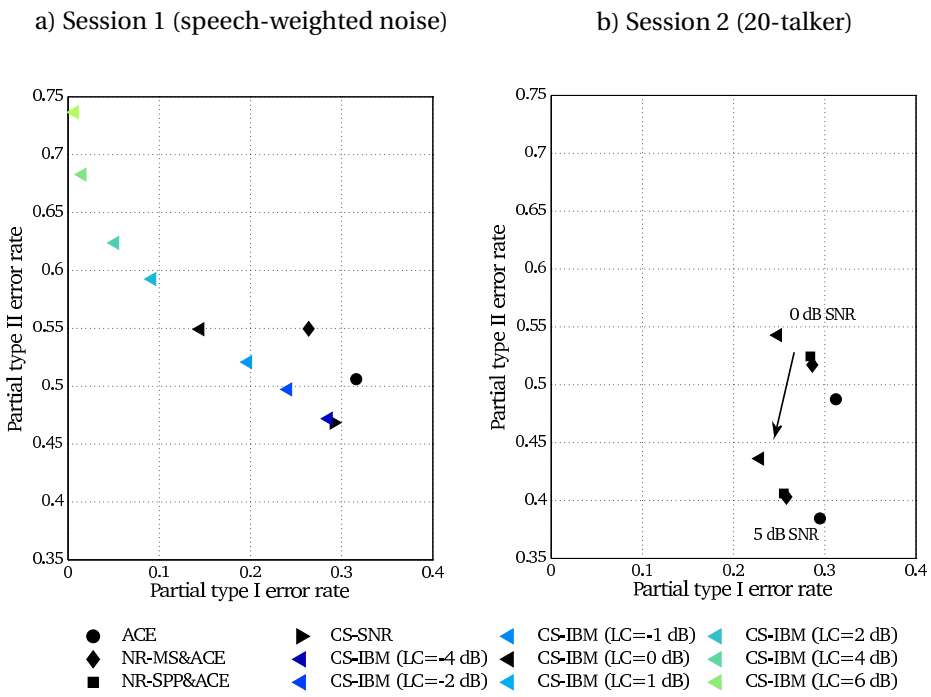


Figure C.1: Partial type I versus type II error rates. In Fig. C.1a, error rates were shown for strategies in session 1 using speech-weighted noise at 0 dB SNR. The "CS-IBM" strategy was simulated for an LC between −4 dB and 6 dB. In Fig. C.1b, error rates were shown for strategies in session 2 using 20-talker noise in 5 and 0 dB SNR, respectively. Partial error components were computed as the difference in intensity level between the estimated and the ideal electrodogram (Hersbach, 2014) over electrodograms from 9 concatenated sentences. A positive difference in intensity was considered a partial type I error component and a negative difference in intensity a partial type II error component for a given CI channel and stimulation cycle. If the difference exceeded 40 dB in dynamic range, the error component was considered a full type I or type II error component.

positive difference in intensity was considered a partial type I error component and a negative difference in intensity a partial type II error component. If the difference exceeded 40 dB in dynamic range, the error component was considered a full type I or type II error component. A full type I error component (a false alarm) was defined as stimulation in a specific CI channel in the estimated electrodogram and no stimulation in the ideal electrodogram. A full type II error component (a miss) was defined as stimulation in a specific CI channel in the ideal electrodogram and no stimulation in the estimated electrodogram. All the type I and type II error components were then added across the electrodogram and converted into partial error rates (Hersbach, 2014).

Figure C.1 shows the partial type I versus partial type II error rates with the strategies in session 1 (Fig. C.1a) and session 2 (Fig. C.1b). In Fig C.1a, the local criterion (LC) was changed between −4 dB and 6 dB in the "CS-IBM" strategy to selecting an appropriate LC for the listener study. A higher LC decreased the partial type I error rate; however, at the expense of an increased partial type II error rate because fewer channels were stimulated per cycle. An LC of 0 dB was selected. This LC was a trade-off between the two error rates and has previously been used (Hu and Loizou, 2008). The "CS-IBM" strategy with an LC of 0 dB had a much lower partial type I error rate and approximately the same partial type II error rate than with the "NR-MS&ACE" strategy. This indicated that the "CS-IBM" strategy missed the same amount of speech per stimulation cycle as the "NR-MS&ACE" strategy; however, with fewer false alarms. Furthermore, lower partial type I and II error rates were found with the "CS-SNR" strategy than with the "ACE" strategy (Fig. C.1a).

In Fig. C.1b, the partial error rates were shown for both 0 dB and for 5 dB SNR in 20-talker noise. It was observed that the "CS-IBM" increased the number of misses over both the "ACE" and the "NR-MS&ACE" strategies; however, with less false alarms. In addition, the "NR-SPP&ACE" and the "NR-MS&ACE" strategies led to approximately the same partial error rates. Therefore, the partial error analysis suggested no differences in performance between these two strategies.

C.3 An evaluation of the speech coding strategies in quiet

Session 2 also tested monosyllabic word recognition in quiet using consonant–vowel nucleus–consonant (CNC) words, to compare with advanced combination encoder (ACE). The purpose was to evaluate any potential degradation in word perception in quiet with the speech coding strategies. One list of 50 words was presented at 65 dB SPL. Results did not show any significant degradation in speech in quiet when comparing each of the strategies with "ACE". Furthermore, the "NR-SPP&ACE" strategy actually led to statistically higher CNC scores than the "NR-MS&ACE" by 5.20% ($p < 0.05$), which implied that this strategy performed better in speech in quiet.

The end.

To be continued...

Understanding speech in noise can be challenging for many people, in particular hearing-aid users and cochlear-implant recipients. To improve the speech understanding, better noise-reduction strategies are needed in such devices. The performance of the strategies depends on how well the characteristics of the speech and the noise are known. Therefore, it is necessary to have automatic approaches that can separate the speech from the noise as accurate as possible, which is the overall goal of computational speech segregation. Often, an ideal time-frequency mask is estimated by auditory-inspired feature extraction combined with machine-learning techniques. In the mask, the level of speech activity is indicated in each unit. This thesis investigated three approaches within computational speech segregation, based on ideal time-frequency mask estimation, and evaluated the approaches in the framework of noise reduction to improve speech understanding of normal-hearing listeners and cochlear-implant recipients in noisy environments. Specifically, the following components were investigated: the spectro-temporal contextual information in speech, the machine-learning system architecture and the ideal time-frequency mask as a learning objective in computational speech segregation. In addition, a practical application of the estimated time-frequency mask was considered in real-time cochlear-implant processing. Overall, the results of this thesis have implications for the design of computational speech segregation approaches with noise-reduction applications. Furthermore, the results may guide the development of a single cost function, which correlates with speech intelligibility, to assess and optimize the system performance.

DTU Electrical Engineering
Department of Electrical Engineering

Ørsteds Plads
Building 348
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 38 00
Fax: (+45) 45 93 16 34
www.elektro.dtu.dk